Journal of Machine Learning Research 24 (2023) 1-34 Submitted 6/22; Published 7/23

Single Timescale Actor-Critic Method to Solve the Linear Quadratic Regulator with Convergence Guarantees

Mo Zhou MO.ZHOU366@DUKE.EDU
Department of Mathematics
Duke University
Durham, NC 27708, USA

Jianfeng Lu JIANFENG@MATH.DUKE.EDU
Department of Mathematics, Department of Physics, and Department of Chemistry
Duke University
Durham, NC 27708, USA

Editor: John Shawe-Taylor

Abstract

We propose a single timescale actor-critic algorithm to solve the linear quadratic regulator (LQR) problem. A least squares temporal difference (LSTD) method is applied to the critic and a natural policy gradient method is used for the actor. We give a proof of convergence with sample complexity $\mathcal{O}(\varepsilon^{-1}\log(\varepsilon^{-1})^2)$. The method in the proof is applicable to general single timescale bilevel optimization problems. We also numerically validate our theoretical results on the convergence.

Keywords: linear quadratic regulator, actor-critic, reinforcement learning, single timescale

1. Introduction

Reinforcement learning (RL) is a semi-supervised learning model that learns to take actions and interact with the environment in order to maximize the expected reward (Sutton and Barto, 2018). It has a wide range of applications, including robotics (Kober et al., 2013), traditional games (Silver et al., 2016), and traffic light control (Wiering, 2000). RL is closely related to the optimal control problem (Bertsekas, 2019), where one usually minimizes the expected cost instead of maximizing the reward. Among all the control problems, the LQR (Anderson and Moore, 2007) is the cleanest setup to analyze theoretically and has many applications (Hashim, 2019; Ebrahim et al., 2010). Many research has been devoted to LQR. Early research mostly focused on model-based methods, such as deriving the explicit solution of the LQR with known dynamics. This research showed that the optimal control is a linear function of the state and the coefficient can be obtained by solving the Riccati equation (Anderson and Moore, 2007). Recent research focuses more on the model-free setting in the context of RL, where the algorithm does not know the dynamic and has only observations of states and rewards (Tu and Recht, 2018; Mohammadi et al., 2021).

The actor-critic method (Konda and Tsitsiklis, 2000) is a class of algorithms that solve the RL or optimal control problems through alternately updating the actor and the critic. In this framework, we solve for both the control and the value function, which is the expected cost w.r.t. the initial state (and action). The control is known as the actor, so in the actor update, we improve the control in order to minimize the cost; i.e., policy improvement. The value function is known as the critic.

Hence, in the critic update, we evaluate a fixed control through computing the value function; i.e., policy evaluation.

On a broader scale, the actor-critic method belongs to the bilevel optimization problem (Sinha et al., 2017; Bard, 2013), as it is an optimization problem (higher-level problem) whose constraint is another optimization problem (lower-level problem). In the actor-critic method, the higher-level problem is to minimize the cost (the actor) and the lower-level problem is to let the critic be equal to value function corresponding to the control, which is equivalent to minimizing the expected squared Bellman residual (Bradtke and Barto, 1996). The major difficulty of a bilevel optimization problem is that when the lower-level problem is not solved exactly, the error could propagate to the higher-level problem and accumulate in the algorithm. One approach to overcome this problem is the two timescale method (Konda and Tsitsiklis, 2000; Wu et al., 2020; Zeng et al., 2021), where the update of lower-level problem is in a time scale that is much faster than the higher-level one. This method suffers from high computational costs because of the lower-level optimization. Another method is to modify the update direction to improve accuracy (Kakade, 2001), which also introduces extra cost. In order to reduce the cost, we seek an efficient single timescale method to solve LQR.

1.1 Our contributions

In this paper, we consider a single timescale actor-critic algorithm to solve the LQR problem. We apply an LSTD method (Bradtke and Barto, 1996) for the critic and a natural policy gradient method (Kakade, 2001) for the actor. For the critic, we derive an explicit expression for the gradient and design a sample method with the desired accuracy, with access to multiple next-step samples from a state. For the actor, we apply a natural policy gradient method borrowed from Fazel et al. (2018). We give a proof of convergence with sample complexity $\mathcal{O}(\varepsilon^{-1}\log(\varepsilon^{-1})^2)$ to achieve an ε-optimal solution. The major challenge is to analyze the interdependent actor and critic part of the algorithm and give bounds for the errors. To the best of our knowledge, our work is the first single timescale actor-critic method to solve the LQR problem with provable guarantees.

Our work not only solves the specific LQR problem but also advances the study of convergence for single timescale bilevel optimization. In our proof of convergence, we construct a Lyapunov function that involves both the critic error and the actor loss. We show that there is a contraction of the Lyapunov function in the algorithm. If we consider the actor and the critic separately, the critic error becomes an issue when we want to show an improvement of the actor and vice versa. Therefore, the higher and lower level problems have to be analyzed simultaneously for a single timescale algorithm.

1.2 Related works

Let us compare our work with related ones in the literature. Perhaps the most closely related work to ours is by Fu et al. (2020). They consider a single timescale actor-critic method to solve the optimal control problem with discrete state and action spaces, while we solve the LQR problem with continuous state and action spaces. They add an entropy regularization in the loss function and achieve a sample complexity of $\mathcal{O}(\varepsilon^{-2})$ with linear parameterization.

For two timescale approaches, Yang et al. (2019) study a two timescale actor-critic algorithm to solve the LQR problem in continuous space. They also use a natural policy gradient method for the actor (Fazel et al., 2018). For the critic, they reformulate policy evaluation into a minimax optimization problem using Fenchel's duality. Several critic steps are performed between two actor

steps and their final sample complexity is $\mathcal{O}(\varepsilon^{-5})$. Zeng et al. (2021) study a bilevel optimization problem that is applied to a two timescale actor-critic algorithm on LQR. They obtain a complexity of $\mathcal{O}(\varepsilon^{-3/2})$. They have assumed strong convexity of the higher-level loss function (actor) while our analysis does not require such assumptions.

Besides model-free approaches, another way to solve the LQR problem is to first learn the model through the system identification approach and then solve the model-based LQR. For example, Dean et al. (2020) use a least square system identification approach to learn the model parameter and then solve the LQR, with sample complexity $\mathcal{O}(\varepsilon^{-2})$. Their work is further improved by Mania et al. (2019), who study the certainty equivalent controller on LQ problem, under both fully observed and partially observed settings.

As can be seen from the above discussions, our single timescale algorithm achieves a lower sample complexity $\mathcal{O}(\varepsilon^{-1}\log(\varepsilon^{-1})^2)$, which is an improvement over previously proposed algorithms.

For the general bilevel optimization problem, we refer the reader to Chen et al. (2022), where the authors summarize the existing bilevel algorithms and propose a STABLE method with $\mathcal{O}(\varepsilon^{-1})$ sample complexity under strong convexity assumption.

The rest of this paper is organized as follows. In Section 2, we introduce the theoretical background of the LQR problem. In Section 3, we describe the algorithm for the LQR problem and our choice of parameters. In Section 4, we give the outline of the convergence proof of the algorithm, with proof details in the appendix. The numerical examples are also deferred to the appendix.

2. Theoretical background

First, we clarify some notations. We use $\|\cdot\|$ to denote the operator norm of a matrix and $\|\cdot\|_F$ to denote the Frobenius norm of a matrix. When we write $M \geq c$ where M is a symmetric matrix and c is a number, we mean $M - cI$ is positive semi-definite. Similarly, $M > c$ means $M - cI$ is positive definite.

We consider a discrete-time Markov process $\{x_s\}$ on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_s\}, \mathbb{P})$:

$$x_{s+1} = Ax_s + Bu_s + \xi_s,$$

where $x_s \in \mathbb{R}^d$ is an adapted state process, $u_s \in \mathbb{R}^k$ is the adapted control process, $A \in \mathbb{R}^{d\times d}$ and $B \in \mathbb{R}^{d\times k}$ are two fixed matrices. $\xi_s \sim N(0, D_\xi)$ is independent noise. The initial state x_0 follows certain distribution that will be specified later.

The goal is to minimize the infinite horizon cost functional

$$J(\{u_s\}) = \lim_{S\to\infty} \mathbb{E}\left[\frac{1}{S}\sum_{s=0}^{S-1} c(x_s, u_s)\right], \tag{1}$$

where $c(x,u) = x^\top Qx + u^\top Ru$ is the one-step cost, with $Q \in \mathbb{R}^{d\times d}$ and $R \in \mathbb{R}^{k\times k}$ being positive definite. Theoretical results guarantee that the optimal control u^* is linear in x: $u_s^* = -K^*x_s$. If the model is known, we can obtain the optimal control parameter by $K^* = (R + B^\top P^* B)^{-1}B^\top P^* A$ where P^* is the solution to the Riccati equation (Anderson and Moore, 2007)

$$P^* = Q + A^\top P^* A - A^\top P^* B(R + B^\top P^* B)^{-1}B^\top P^* A. \tag{2}$$

In this work, we consider the model-free setting (i.e., the algorithm does not have access to A, B, D_ξ, Q, R). We will use a stochastic policy parametrized as

$$u_s \sim \pi_K := N(-Kx_s, \sigma^2 I_k) \tag{3}$$

to encourage exploration, where $\sigma > 0$ is a fixed constant. Here, we use π_K to denote the distribution while we will not distinguish in notation a probability distribution with its density. We remark that adding exploration does not change the optimal K^* because the optimal policy parameters with or without exploration satisfy the same Riccati equation while adding exploration would help the convergence of the algorithm. Under this policy, the cost functional (1) is also denoted by $J(K)$ and the state trajectory can be rewritten as

$$x_{s+1} = Ax_s + B(-Kx_s + \sigma\omega_s) + \xi_s =: (A - BK)x_s + \epsilon_s$$

where $\omega_s \sim N(0, I_k)$ and $\epsilon_s \sim N(0, D_\epsilon)$ with $D_\epsilon = D_\xi + \sigma^2 BB^\top$ being positive definite. Let $\rho(\cdot)$ denote the spectral radius of a matrix. When $\rho(A - BK) < 1$, the state process has a stationary distribution $N(0, D_K)$, where $D_K \in \mathbb{R}^{d\times d}$ satisfies the Lyapunov equation

$$D_K = D_\epsilon + (A - BK)D_K(A - BK)^\top. \tag{4}$$

In order to understand (4), let us assume that $x \sim N(0, D_K)$ follows the stationary distribution. Then, $x' = (A - BK)x + \epsilon \sim N(0, (A - BK)D_K(A - BK)^\top + D_\epsilon)$ also follows the stationary distribution, which leads to (4). D_K can also be expressed in terms of a series: since $\rho(A - BK) < 1$, we can recursively plug in the definition of D_K into the right hand side of (4) and obtain

$$D_K = \sum_{s=0}^{\infty} (A - BK)^s D_\epsilon ((A - BK)^\top)^s. \tag{5}$$

From here on, the notation $\mathbb{E}_K$ means the expectation with x (or x_0) $\sim N(0, D_K)$ if not specified and u (or u_s) $\sim \pi_K$. The state-action value function (Q function) and the state value function with respect to a control $\{u_s\}$ are defined by

$$\begin{aligned} Q(x, u) &= \sum_{s=0}^{\infty} \left(\mathbb{E}\left[c(x_s, u_s) \mid x_0 = x, u_0 = u\right] - J(\{u_s\})\right) \\ V(x) &= \sum_{s=0}^{\infty} \left(\mathbb{E}\left[c(x_s, u_s) \mid x_0 = x\right] - J(\{u_s\})\right) = \mathbb{E}_u\left[Q(x, u)\right] \end{aligned} \tag{6}$$

respectively. $V(x)$ is the expected extra cost if we start at $x_0 = x$ and follow a given policy. $Q(x, u)$ is the expected extra cost if we start at $x_0 = x$, take the first action $u_0 = u$, and then follow a given policy. These two functions are crucial in reinforcement learning. If the policy π_K follows (3), then the two functions in (6) are denoted by $Q_K(x, u)$ and $V_K(x)$ respectively. By definition, for any x and u, it satisfies the Bellman equation:

$$Q_K(x, u) = c(x, u) - J(K) + \mathbb{E}_K\big[Q_K(x', u') \mid x, u\big], \tag{7}$$

where (x', u') is the next state-action pair starting from (x, u).

We define P_K as the solution to the following matrix valued equation

$$P_K = (Q + K^\top RK) + (A - BK)^\top P_K(A - BK). \tag{8}$$

P_K can be interpreted as the second order adjoint state, and $P_K x_t$ is the shadow price for the system (see for example Yong and Zhou (1999)). We have the following two properties to illustrate the importance of P_K. The proofs are deferred to the appendix.

Proposition 1 *Let the policy π_K be defined by* (3) *with $\rho(A - BK) < 1$. Then the cost function and its gradient w.r.t. K have the following explicit expressions:*

$$J(K) = \operatorname{Tr}(D_\epsilon P_K) + \sigma^2 \operatorname{Tr}(R), \tag{9}$$

$$\nabla_K J(K) = 2\left[(R + B^\top P_K B)K - B^\top P_K A\right] D_K. \tag{10}$$

Remark 1 *In the LQR problem, we usually assume that D_K is positive definite and hence invertible. Therefore, the critical point for $J(K)$ (i.e., when $\nabla_K J(K) = 0$) satisfies $K = (R + B^\top P_K B)^{-1} B^\top P_K A$. If we substitute this into* (8)*, we recover the Riccati equation* (2)*.*

Proposition 2 *Let the policy π_K be defined by* (3) *with $\rho(A - BK) < 1$. Then the value functions have the following explicit expressions:*

$$V_K(x) = x^\top P_K x - \operatorname{Tr}(D_K P_K),$$

$$Q_K(x,u) = \begin{bmatrix} x^\top & u^\top \end{bmatrix} \begin{bmatrix} Q + A^\top P_K A & A^\top P_K B \\ B^\top P_K A & R + B^\top P_K B \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} - \sigma^2 \operatorname{Tr}(R + P_K B B^\top) - \operatorname{Tr}(D_K P_K). \tag{11}$$

If we concatenate x and u in the dynamic equation, the process can be written as

$$\begin{bmatrix} x_{s+1} \\ u_{s+1} \end{bmatrix} = \begin{bmatrix} A & B \\ -KA & -KB \end{bmatrix} \begin{bmatrix} x_s \\ u_s \end{bmatrix} + \begin{bmatrix} \xi_s \\ -K\xi_s + \sigma\omega_s \end{bmatrix}.$$

We simplify the expression by introducing some new notations: $z_s = [x_s^\top, u_s^\top]^\top$, thus $z_{s+1} = Ez_s + \widetilde{\epsilon}_s$, where

$$E = \begin{bmatrix} A & B \\ -KA & -KB \end{bmatrix}, \text{ and } \widetilde{\epsilon}_s \sim N(0, \Sigma_\epsilon) := N\left(0, \begin{bmatrix} D_\xi & -D_\xi K^\top \\ -KD_\xi & KD_\xi K^\top + \sigma^2 I_k \end{bmatrix}\right). \tag{12}$$

The ergodicity of the dynamics is guaranteed if $\rho(A - BK) = \rho(E) < 1$, where the identity $\rho(A - BK) = \rho(E)$ can be verified from

$$\rho(E) = \rho\left(\begin{bmatrix} I_d \\ -K \end{bmatrix} \begin{bmatrix} A & B \end{bmatrix}\right) = \rho\left(\begin{bmatrix} A & B \end{bmatrix} \begin{bmatrix} I_d \\ -K \end{bmatrix}\right) = \rho(A - BK).$$

The stationary distribution for z is given by

$$z \sim N(0, \Sigma_K) := N\left(0, \begin{bmatrix} D_K & -D_K K^\top \\ -KD_K & KD_K K^\top + \sigma^2 I_k \end{bmatrix}\right) \tag{13}$$

and we have $\Sigma_K = \Sigma_\epsilon + E\Sigma_K E^\top$.

3. The actor-critic algorithm

In this section, we present our specific design of the algorithm under the actor-critic framework. We apply an LSTD method for the policy evaluation (critic), with a detailed description for sampling the gradient of the loss function. We also use a natural policy gradient method for the policy improvement (actor). We will use $\mathcal{G}_t$ to denote the filtration generated by the training process. We use $\mathcal{O}(a)$ to denote a quantity that is is bounded by a constant times a, where this constant only depends on the problem setting (A, B, D_ϵ, Q, R, σ) and does not depend on the target accuracy or training trajectory. The dependence of the constants on the dimensions is explained in the proof of our theorem.

3.1 Policy evaluation for the critic

In this subsection, we describe the policy evaluation algorithm for a fixed policy π_K. We parametrize the state-action value function by Q_K^θ with θ as a parameter and subscript K indicating that it depends on the given policy π_K. We define the Bellman residual w.r.t. the critic parameter θ as

$$\mathsf{BR}_\theta(x,u) = c(x,u) - J(K) + \mathbb{E}_K\left[Q_K^\theta(x',u')|x,u\right] - Q_K^\theta(x,u).$$

Recall the exact Q function is given by (11), accordingly, we define a feature matrix

$$\phi(x,u) = \begin{bmatrix} x \\ u \end{bmatrix}\begin{bmatrix} x^\top & u^\top \end{bmatrix} \in \mathbb{R}^{(d+k)\times(d+k)} \tag{14}$$

and parametrize the Q function as

$$Q_K^\theta(x,u) = \mathrm{Tr}(\phi(x,u)\theta) - \theta', \tag{15}$$

where $\theta \in \mathbb{R}^{(d+k)\times(d+k)}$ and $\theta' \in \mathbb{R}$. Here, we denote

$$\theta = \begin{bmatrix} \theta^{11} & \theta^{12} \\ \theta^{21} & \theta^{22} \end{bmatrix}, \text{ which intends to approximate } \theta_K = \begin{bmatrix} Q + A^\top P_K A & A^\top P_K B \\ B^\top P_K A & R + B^\top P_K B \end{bmatrix}. \tag{16}$$

The scalar parameter θ' is to approximate $\sigma^2\,\mathrm{Tr}(R + P_K B B^\top) + \mathrm{Tr}(D_K P_K)$. Recall the Bellman equation (7), with parametrization (15), the Bellman residual is written as

$$\begin{aligned}\mathrm{BR}_\theta(x,u) = c(x,u) - J(K) + \langle \mathbb{E}_K\left[\phi(x',u')|x,u\right] - \phi(x,u), \theta\rangle \\ =: c(x,u) - J(K) + \langle \psi(x,u), \theta\rangle,\end{aligned}$$

where $\langle\cdot,\cdot\rangle$ is the trace inner product and we have defined $\psi(x,u) := \mathbb{E}_K\left[\phi(x',u')|x,u\right] - \phi(x,u)$ for convenience. It is clear by definition that $\mathbb{E}_K[\psi(x,u)] = 0$ (recall that x follows the stationary distribution $N(0, D_K)$). The loss function for critic is then defined as the expectation of squared Bellman residual:

$$L_K(\theta) = \frac{1}{2}\mathbb{E}_K\left[\mathrm{BR}_\theta(x,u)^2\right] = \frac{1}{2}\mathbb{E}_K\left[\left(c(x,u) - J(K) + \langle\psi(x,u),\theta\rangle\right)^2\right]. \tag{17}$$

We will find that θ' does not affect the training, so only θ will be considered as the critic parameter from now on. According to the Bellman equation (7), the unique minimizer of (17) is the true

parameter for the Q function w.r.t. π_K. By direct computation, the gradient (as a matrix) and Hessian (as a tensor) of the loss function w.r.t. θ are

$$\begin{aligned}\nabla L_K(\theta) &= \mathbb{E}_K\left[\left(c(x,u) - J(K) + \langle \psi(x,u), \theta\rangle\right)\psi(x,u)\right] \\ &= \mathbb{E}_K\left[\left(c(x,u) + \langle \psi(x,u), \theta\rangle\right)\psi(x,u)\right]\end{aligned} \tag{18}$$

and

$$\nabla^2 L_K(\theta) = \mathbb{E}_K\left[\psi(x,u)\otimes\psi(x,u)\right],$$

where $\otimes$ denotes the tensor product. The loss function L_K is strongly convex in θ, as will be shown later.

To minimize the loss (17), we use stochastic gradient descent method. Thus, we need an accurate sample estimate of $\nabla L_K(\theta)$ for given K and θ. For simplicity of notation, we denote

$$f(x,u) := \left(c(x,u) + \langle \psi(x,u), \theta\rangle\right)\psi(x,u) = c(x,u)\psi(x,u) + \left(\psi(x,u)\otimes\psi(x,u)\right)\cdot\theta \tag{19}$$

so that $\nabla L_K(\theta) = \mathbb{E}_K[f(x,u)]$. Note that $f(x,u)$ depends on θ and K, while we suppress that in the notation. We decompose the sampling into three steps: we firstly sample $\psi(x,u)$, then sample $f(x,u)$ accordingly, and finally give estimate of $\nabla L_K(\theta)$.

For the first step, we use the Markov chain Monte Carlo (MCMC) method (Gilks et al., 1995). Let N_0 and N be two integers that will be determined according to the error tolerance. Starting at $x_0 = 0$, we sample N independent trajectories of length N_0+1 according to the policy π_K. So, we obtain N samples $\{(x_{N_0}^{(i)}, u_{N_0}^{(i)})\}_{i=1}^N$ that follow the distribution of (x_{N_0}, u_{N_0}). For each pair $(x_{N_0}^{(i)}, u_{N_0}^{(i)})$, we generate N_1 unbiased sample for $\psi(x_{N_0}^{(i)}, u_{N_0}^{(i)})$, given by

$$\widehat{\psi}_j^{(i)} = \phi(x^{(i,j)}, u^{(i,j)}) - \phi(x_{N_0}^{(i)}, u_{N_0}^{(i)}) \qquad j = 1, 2, \cdots, N_1$$

where $x^{(i,j)}, u^{(i,j)}$ are sampled independently and follow the next step distribution conditioned on $(x_{N_0}^{(i)}, u_{N_0}^{(i)})$. Here, N_1 is another predefined hyperparameter.

In the second step, we denote the mean of $\widehat{\psi}_j^{(i)}$ by $\bar{\psi}^{(i)} = \frac{1}{N_1}\sum_{j=1}^{N_1}\widehat{\psi}_j^{(i)}$. Therefore, we can obtain an unbiased sample for $f(x_{N_0}^{(i)}, u_{N_0}^{(i)})$ by

$$\begin{aligned}\widehat{f}(x_{N_0}^{(i)}, u_{N_0}^{(i)}) =& \frac{1}{N_1}\sum_{j=1}^{N_1} c(x_{N_0}^{(i)}, u_{N_0}^{(i)})\widehat{\psi}_j^{(i)} \\ &+ \left[\frac{1}{N_1}\sum_{j=1}^{N_1}\widehat{\psi}_j^{(i)}\otimes\widehat{\psi}_j^{(i)} - \frac{1}{N_1-1}\sum_{j=1}^{N_1}(\widehat{\psi}_j^{(i)} - \bar{\psi}^{(i)})\otimes(\widehat{\psi}_j^{(i)} - \bar{\psi}^{(i)})\right]\cdot\theta.\end{aligned} \tag{20}$$

Note that the first and second terms in the square bracket are unbiased samples for $\mathbb{E}[\widehat{\psi}_j^{(i)}\otimes\widehat{\psi}_j^{(i)}]$ and $\mathrm{Cov}(\widehat{\psi}_j^{(i)})$ respectively, which implies that the square bracket is an unbiased sample for $\psi(x_{N_0}^{(i)}, u_{N_0}^{(i)})\otimes\psi(x_{N_0}^{(i)}, u_{N_0}^{(i)})$. Note that we require $N_1 \geq 2$, which implies our algorithm is not a pure online method.

Finally, the sample of gradient $\nabla L_K(\theta)$ is given by

$$\widehat{\nabla L}_K(\theta) = \frac{1}{N}\sum_{i=1}^{N}\widehat{f}(x_{N_0}^{(i)}, u_{N_0}^{(i)}). \tag{21}$$

The one-step sample complexity is $\mathcal{O}(N_0 N_1 N)$. We remark that our LSTD is similar to a TD(0) algorithm, except that we have N trajectories and we omit $J(K)$ in (18). Denote $L_{K_t}(\theta)$ by $L_t(\theta)$ for simplicity. We also denote θ_t the critic parameter at step t. The gradient sample at step t (in matrix form) is denoted by $\widehat{\nabla L_t}(\theta_t)$ and the critic update is given by

$$\theta_{t+1} = \theta_t - \alpha_t \widehat{\nabla L_t}(\theta_t),$$

where α_t is the step size for the critic.

3.2 Policy improvement for the actor

For the actor algorithm, we borrow the idea from Fazel et al. (2018) which considered a policy gradient algorithm for the LQR problem. A similar approach is also studied by Yang et al. (2019); Zeng et al. (2021).

Motivated by the form of the gradient (10), we define

$$G_K := (R + B^\top P_K B)K - B^\top P_K A, \tag{22}$$

so that $\nabla_K J(K) = 2G_K D_K$. Therefore, a vanilla policy gradient algorithm looks like

$$K_{t+1} = K_t - \beta_t G_{K_t} D_{K_t},$$

where G_{K_t} and D_{K_t} may be replaced by some estimates and β_t is the step size for the actor.

Instead of the vanilla policy gradient, we would consider the commonly used variant known as the natural policy gradient method (Kakade, 2001). The natural policy gradient uses the inverse Fisher information matrix to precondition the gradient so that the gradient is taken w.r.t. the metric induced by the Hessian of the loss function (Peters and Schaal, 2008). This method has been studied in e.g., (Kakade, 2001; Peters and Schaal, 2008; Bhatnagar et al., 2009; Liu et al., 2020). The Fisher information matrix at each state x is given by

$$F_x(K) = \mathbb{E}_{u\sim\pi_K}\left[\nabla_K \log(\pi_K(u|x)) \otimes \nabla_K \log(\pi_K(u|x))\right], \tag{23}$$

which is a tensor in $\mathbb{R}^{k\times d} \otimes \mathbb{R}^{k\times d}$ as $K \in \mathbb{R}^{k\times d}$ is a matrix. Then, the (average) Fisher information matrix is defined as

$$F(K) = \mathbb{E}_{x\sim N(0,D_K)}\left[F_x(K)\right] = \mathbb{E}_K\left[\nabla_K \log(\pi_K(u|x)) \otimes \nabla_K \log(\pi_K(u|x))\right].$$

Under the metric induced by the Hessian, the steepest descent direction of $J(K)$ is given by

$$-\widetilde{\nabla} J(K) = -F(K)^{-1}\nabla_K J(K) = -2F(K)^{-1} G_K D_K,$$

where for $F(K)^{-1}$, we view the tensor $F(K)$ as a linear operator $\mathbb{R}^{k\times d} \to \mathbb{R}^{k\times d}$, so $F(K)^{-1}$ is the inverse operator. The following property gives a simple expression of $\widetilde{\nabla} J(K)$. The proof is in the appendix.

Proposition 3 *We have*

$$\widetilde{\nabla} J(K) = 2\sigma^2 G_K. \tag{24}$$

Recall that $G_K = (R + B^\top P_K B)K - B^\top P_K A$. Hence, $G_K = \theta_K^{22} K - \theta_K^{21}$ where θ_K is the true parameter w.r.t. policy π_K, given by (16). Therefore, the actor update is given by

$$K_{t+1} = K_t - \beta_t(\theta_t^{22} K_t - \theta_t^{21}) =: K_t - \beta_t \widehat{G}_{K_t}, \tag{25}$$

where the constant $2\sigma^2$ is absorbed in the step size β_t and we have defined $\widehat{G}_{K_t} := \theta_t^{22} K_t - \theta_t^{21}$. Recall that we use $\mathcal{G}_t$ to denote the filtration generated by the training process. Since K_{t+1} is deterministic in θ_t and K_t, K_{t+1} is $\mathcal{G}_t$-measurable.

3.3 Assumptions and main result

Here we state some technical assumptions for our result.

Assumption 1 *We assume that*

1. *There exists a constant $\rho \in (0,1)$ such that $\rho(A - BK_t) = \rho(E_t) \leq \rho$, for all t.*
2. *There exist constants $c_A, c_E, c_\theta, c_K > 0$ such that $\|A - BK_t\| \leq c_A$, $\|E_t\| \leq c_E$, $\|\theta_t\|_F \leq c_\theta$, and $\|K^*\|, \|K_t\| \leq c_K$ for all t.*
3. *D_ϵ is positive definite with minimum eigenvalue $\sigma_{min}(D_\epsilon) > 0$.*

Remark 2 *In the assumption, E_t is defined by* (12) *with K replaced by K_t. The first assumption is common in the analysis of the LQR problem (Fazel et al., 2018; Yang et al., 2019). A theoretical guarantee for this condition is hard to obtain, while we will present some numerical examples to support this assumption. The second assumption gives upper bounds for several matrices, which is made to avoid technical tedious works to control the probability of the random trajectory hitting unfavorable regions. One potential way to alleviate this assumption is to define a projection map that reduces the size of θ_t or K_t whenever it is out of range (Konda and Tsitsiklis, 2000; Bhatnagar et al., 2009), which is left for future work. The third assumption is necessary to make the problem non-degenerate (cf. Lemma 7 below).*

Next, we specify the choice of parameters in the algorithm. We initialize $\theta_0 = 0$, $K_0 = 0$ for simplicity. Fixing the error tolerance $\varepsilon > 0$, we set the step sizes α_t and β_t to be constant in t:

$$\alpha_t = \frac{\sigma_{min}(D_\epsilon)}{16 c_L^2 c_3 \kappa} \varepsilon \qquad \beta_t = \frac{\sigma_{min}(D_\epsilon)}{16 c_L^2 c_3 \kappa^2} \varepsilon \tag{26}$$

where

$$\kappa = \max\left(\frac{3\sigma_{min}(D_\epsilon)}{2 c_3 \mu_\sigma}, \frac{4 c_1^2}{\mu_\sigma \sigma_{min}(D_\epsilon)}, \frac{3 c_D c_K^2}{\mu_\sigma} \right). \tag{27}$$

Here, every parameter appearing in (26) and (27), except α_t, β_t, or ε, are constants of order $\mathcal{O}(1)$:

1. c_L^2 is the upper bound for $\mathbb{E}[\|\widehat{\nabla L_t}(\theta_t)\|_F^2 \mid \mathcal{G}_t]$ that is in Lemma 3;
2. c_3 illustrates the geometry of $J(K)$, with details in Lemma 6;
3. In Lemma 2, we will show that the critic loss is μ_σ-strongly convex;
4. c_1 is a Lipschitz constant for θ_K w.r.t. K that is specified in Lemma 4;
5. c_D is an upper bound for $\|D_{K_t}\|$ and $\|D_{K^*}\|$ that is specified in Lemma 1.

It is easy to verify that the step sizes satisfies the following inequalities:

$$\frac{\sigma_{min}(D_\epsilon)}{c_3}\beta_t \leq \frac{2}{3}\mu_\sigma \alpha_t, \ \frac{\sigma_{min}(D_\epsilon)}{\beta_t} \geq (\frac{3}{\alpha_t \mu_\sigma} + 2)c_1^2 + (\|R\| + c_P\|B\|^2), \text{ and } \frac{1}{3}\alpha_t \mu_\sigma \geq \beta_t c_D c_K^2, \tag{28}$$

where we need to assume that ε is small enough such that $1/(\mu_\sigma \alpha_t) \geq 2 + (\|R\| + c_P\|B\|^2)/c_1^2$ for the second inequality. Here, c_P is the upper bound for P_{K_t}, which is given in Lemma 1. These are technical inequalities that will be used in the proof later. The total number of iterations is $T = \mathcal{O}(\frac{1}{\varepsilon}\log(\frac{1}{\varepsilon}))$ such that

$$(1 - \beta_t c_4)^T L_0 < \varepsilon,$$

where $L_0 = \mathcal{O}(1)$ is the initial Lyapunov function that is specified at the beginning of the proof for Theorem 1 and $c_4 = \mathcal{O}(1)$ is a positive constant that is also specified in the proof for Theorem 1. This $(1 - \beta_t c_4)$ is the one-step decay ratio of the Lyapunov function, which indicates our choice of T above. The number of samples N, the length of trajectory N_0 each step, and the sub-sample size N_1, are set to be $N = \mathcal{O}(1)$, $N_0 = \mathcal{O}(\log(\frac{1}{\varepsilon}))$, and $N_1 = \mathcal{O}(1)$, in order to achieve desired accuracy for the sample of critic gradient, with details in Lemma 3. Here, $\frac{\alpha_t}{\beta_t} = \kappa = \mathcal{O}(1)$ implies that our algorithm has single timescale. In such algorithm, the actor and the critic are interdependent, which makes the analysis challenging. We summarize the actor-critic algorithm in Algorithm 1.

Algorithm 1 Single timescale actor-critic algorithm for LQR

Input: Training steps T, step sizes α_t, β_t, sample size N, N_0, and N_1
Output: critic parameter θ_T, actor parameter K_T
initialization: critic parameter $\theta_0 = 0$ and actor parameter $K_0 = 0$
for $t = 0$ **to** $T - 1$ **do**
Sample $\widehat{\nabla L_t}(\theta_t)$ according to (21) ▷ critic steps
$\theta_{t+1} = \theta_t - \alpha_t \widehat{\nabla L_t}(\theta_t)$
$K_{t+1} = K_t - \beta_t(\theta_t^{22} K_t - \theta_t^{21})$ ▷ actor steps
end for

The main result of our work is the following convergence theorem.

Theorem 1 (Main theorem) *Under Assumption 1, for any $\varepsilon > 0$ that is sufficiently small, Algorithm 1, with the choice of parameters discussed above, has sample complexity $\mathcal{O}(\frac{1}{\varepsilon}\log(\frac{1}{\varepsilon})^2)$. Moreover, the terminal error satisfies*

$$\mathbb{E}[\|\theta_T - \theta_{K_T}\|_F^2] \leq \varepsilon \quad \textit{and} \quad \mathbb{E}[J(K_T) - J(K^*)] \leq \varepsilon.$$

Remark 3 *The number of steps is $T = \mathcal{O}(\frac{1}{\varepsilon}\log(\frac{1}{\varepsilon}))$ and the one-step complexity is $\mathcal{O}(\log(\frac{1}{\varepsilon}))$. Therefore, the total complexity is $\mathcal{O}(\frac{1}{\varepsilon}\log(\frac{1}{\varepsilon})^2)$. This theorem tells us that we have small error for both the critic and the actor. If we want error estimate for $\|K_T - K^*\|_F$ or $\|\theta_T - \theta^*\|_F$, we will need extra assumption such as strong convexity of $J(K)$ in K.*

As a follow up for Remark 2, another potential way to alleviate Assumption 1 is to modify the main theorem in concentration sense (the result holds with high probability), which omits the rare cases.

We believe the complexity $\mathcal{O}(\frac{1}{\varepsilon}\log(\frac{1}{\varepsilon})^2)$ is nearly optimal (up to a log factor). Even for a simple stochastic gradient descent (SGD) algorithm, we need $\mathcal{O}(\varepsilon^{-1})$ sample to achieve ε-optimal solution (Bottou, 2012). The LQR problem is bilevel, with the critic part similar to SGD. Thus, the problem is more complicated than SGD and expects to require higher sample complexity. The convergence rate is also confirmed by the numerical examples below.

4. Proof sketch of the main theorem

In this section, we give a sketch of the proof of Theorem 1 and postpone the details to the appendix. The lemmas used in the proof are stated in the later part of this section.

Proof [Proof Sketch of Theorem 1] First, we show in Lemma 2 that the critic loss is strongly convex. Then, we show in Lemma 3 that we can obtain the sample of gradient with small bias:

$$\left\| \mathbb{E}\left[\widehat{\nabla L_t}(\theta_t) - \nabla L_t(\theta_t) | \mathcal{G}_t \right] \right\|_F \leq \delta$$

With these two lemmas, we show in Lemma 5 that there is an improvement of critic error in each step:

$$\begin{aligned} &\mathbb{E}\left[\|\theta_{t+1} - \theta_{K_{t+1}}\|_F^2 | \mathcal{G}_t \right] - \|\theta_t - \theta_{K_t}\|_F^2 \\ &\qquad \leq -\frac{4}{3}\alpha_t \mu_\sigma \|\theta_t - \theta_{K_t}\|_F^2 + \frac{1}{4}\frac{\sigma_{min}(D_\epsilon)}{c_3}\beta_t \varepsilon + \big(\frac{3}{\alpha_t \mu_\sigma} + 2\big)\|\theta_{K_t} - \theta_{K_{t+1}}\|_F^2. \end{aligned} \tag{29}$$

Here, the term $\frac{1}{4}\frac{\sigma_{min}(D_\epsilon)}{c_3}\beta_t\varepsilon$ comes from the sample error in Lemma 3 and $(\frac{3}{\alpha_t\mu_\sigma} + 2)\|\theta_{K_t} - \theta_{K_{t+1}}\|_F^2$ is due to the actor update. Intuitively, we expect $\|\theta_{t+1} - \theta_{K_t}\|_F$ to be smaller than $\|\theta_t - \theta_{K_t}\|_F$, recall that $\|\theta_t - \theta_{K_t}\|_F$ measures the error of θ_t w.r.t. the current policy parameter K_t, while the last term in (29) takes into account the update of K_t to K_{t+1} in the actor step.

Furthermore, we establish the improvement of the actor in Lemma 7:

$$\begin{aligned} J(K_{t+1}) - J(K_t) &\leq -\beta_t \frac{\sigma_{min}(D_\epsilon)}{c_3}(J(K_t) - J(K^*)) \\ &- \beta_t \left[\sigma_{min}(D_\epsilon) - \beta_t c_D(\|R\| + c_P\|B\|^2)\right] \|\widehat{G}_{K_t}\|_F^2 + \beta_t c_D \|G_{K_t} - \widehat{G}_{K_t}\|_F^2 \end{aligned} \tag{30}$$

where the last term comes from the critic error.

To establish the convergence, we define a Lyapunov function

$$\mathcal{L}_t = \mathcal{L}(\theta_t, K_t) := \|\theta_t - \theta_{K_t}\|_F^2 + J(K_t) - J(K^*),$$

which is the sum of critic and actor errors. Direct computation shows that the last term in (29) can be bounded by the second term in (30) and the last term in (30) can be bounded by $\frac{1}{4}$ of the first term in (29). Therefore, combining (29) and (30), we obtain the decay estimate of the Lyapunov function

$$\mathbb{E}[\mathcal{L}_{t+1} - \mathcal{L}_t] \leq -\mathbb{E}\left[\alpha_t \mu_\sigma \|\theta_t - \theta_{K_t}\|_F^2 + \beta_t \frac{\sigma_{min}(D_\epsilon)}{c_3}(J(K_t) - J(K^*))\right] + \frac{1}{4}\frac{\sigma_{min}(D_\epsilon)}{c_3}\beta_t\varepsilon. \tag{31}$$

Notice that the last term (sample error) in (31) can be bounded by the first term if $\mathbb{E}[\|\theta_t - \theta_{K_t}\|_F^2] \geq \frac{\varepsilon}{2}$ (according to the first inequality of (28)) or by the second term if $\mathbb{E}[J(K_t) - J(K^*)] \geq \frac{\varepsilon}{2}$ and we will obtain a contraction rate for the Lyapunov function:

$$\mathcal{L}_{t+1} - \mathcal{L}_t \leq -\mathcal{O}(\beta_t)\mathcal{L}_t.$$

If both $\mathbb{E}[\|\theta_t - \theta_{K_t}\|_F^2] < \frac{\varepsilon}{2}$ and $\mathbb{E}[J(K_t) - J(K^*)] < \frac{\varepsilon}{2}$, then $\mathbb{E}[\mathcal{L}_t] < \varepsilon$ and we can easily show that $\mathbb{E}[\mathcal{L}_{t+1}]$ is also less than ε. This finishes the proof. ■

In summary, the key point of the proof is that we can bound the positive term in the critic improvement by the negative term in the actor improvement and vice versa. In this way, we obtain a contraction rate of the Lyapunov function.

Before we turn to the analysis of critic and actor parts, we state the following lemma which provides bounds for matrices D_{K_t}, P_{K_t}, and Σ_{K_t}.

Lemma 1 *Under Assumption 1, the matrix D_{K_t}, P_{K_t} and Σ_{K_t} satisfy*

$$\sigma_{min}(D_\epsilon) \le D_{K_t} \le c_D, \quad P_{K_t} \le c_P, \ \text{ and } \ \Sigma_{K_t} \le c_\Sigma \tag{32}$$

where the three constants $c_D, c_P, c_\Sigma = \mathcal{O}(1)$ only depend on A, B, D_ϵ, Q, R, ρ, σ, and c_A. Furthermore, the first inequality also holds with D_{K_t} replaced by D_{K^*}.

4.1 Analysis of the critic part

In this subsection, we analyze the critic part of the algorithm. All the proofs are deferred to the appendix. Let us start with the following lemma, which gives the strong convexity property of the critic loss.

Lemma 2 (Strong convexity of critic loss) *Suppose that $\rho(E) \le \rho < 1$, $L_K(\theta)$ is μ_σ-strongly convex in θ, where $\mu_\sigma > 0$ only depends on A, B, D_ϵ, ρ, σ, c_K, and c_Σ. Moreover, $\mu_\sigma = \mathcal{O}(\sigma^4)$ when σ is small.*

Actually, one technical reason of using a stochastic policy for exploration is to guarantee the strong convexity. The next lemma gives a quantitative description of the accuracy of critic gradient sampling proposed in §3.1.

Lemma 3 (Gradient sample accuracy) *Under Assumption 1, for any $\delta > 0$ that is sufficiently small, let $\widehat{\nabla L_t}(\theta_t)$ be the sample of $\nabla L_t(\theta_t)$ with complexity $N, N_1 = \mathcal{O}(1)$ and $N_0 = \mathcal{O}(\log \frac{1}{\delta})$. Then, we have*

$$\left\| \mathbb{E}\left[\widehat{\nabla L_t}(\theta_t) - \nabla L_t(\theta_t) \,\middle|\, \mathcal{G}_t \right] \right\|_F \le \delta \tag{33}$$

and

$$\mathbb{E}\left[\|\widehat{\nabla L_t}(\theta_t)\|_F^2 \,\middle|\, \mathcal{G}_t \right] \le c_L^2, \tag{34}$$

where $c_L = \mathcal{O}(1)$ is a positive constant that only depends on A, B, D_ϵ, Q, R, σ, c_K, and c_θ.

Remark 4 *When we apply this lemma later, we will set $\delta^2 = \frac{1}{24}\frac{\sigma_{min}(D_\epsilon)}{\kappa c_3}\mu_\sigma \varepsilon$, and thus $\delta = \mathcal{O}(\varepsilon^{\frac{1}{2}})$. By definition of the step sizes* (26)*, we have*

$$2\alpha_t^2 \mathbb{E}\left[\|\widehat{\nabla L_t}(\theta_t)\|_F^2 \,\middle|\, \mathcal{G}_t \right] \le \frac{1}{8}\beta_t \frac{\sigma_{min}(D_\epsilon)}{c_3}\varepsilon. \tag{35}$$

when (34) *holds. This inequality* (35) *will be used later and we can see that the step size has to be of order $\mathcal{O}(\varepsilon)$ to guarantee* (35)*.*

Next, we show a Lipschitz property for θ_K with respect to K.

Lemma 4 *For any two actor parameters K and K' such that $\|K\|, \|K'\| \le c_K$, $\|A - BK\|, \|A - BK'\| \le c_A$, and $\rho(A - BK), \rho(A - BK') \le \rho < 1$, we have*

$$\|\theta_K - \theta_{K'}\|_F \le c_1 \|K - K'\|_F,$$

where the constant $c_1 = \mathcal{O}(1)$ only depends on A, B, R, ρ, c_A, c_K, and c_P.

With the above lemmas, we can establish the improvement by the critic update.

Lemma 5 *Let the step size be defined as in* (26) *and Assumption 1 hold. For any $\varepsilon > 0$ that is sufficiently small, assume that* (33) *and* (34) *hold with $\delta^2 = \frac{1}{24}\frac{\sigma_{min}(D_\epsilon)}{\kappa c_3}\mu_\sigma\varepsilon$ for all t, then we have*

$$\mathbb{E}\left[\|\theta_{t+1} - \theta_{K_{t+1}}\|_F^2 \mid \mathcal{G}_t\right] - \|\theta_t - \theta_{K_t}\|_F^2 \\ \leq -\frac{4}{3}\alpha_t\mu_\sigma\|\theta_t - \theta_{K_t}\|_F^2 + \frac{1}{4}\frac{\sigma_{min}(D_\epsilon)}{c_3}\beta_t\varepsilon + \big(\frac{3}{\alpha_t\mu_\sigma} + 2\big)\|\theta_{K_t} - \theta_{K_{t+1}}\|_F^2. \quad (36)$$

Recall that K_{t+1} is $\mathcal{G}_t$-measurable.

4.2 Analysis of the actor part

In this subsection, we give the convergence result for the actor part. All proofs are deferred to the appendix. The first lemma demonstrates that the cost functional is roughly quadratic in G_K. Inequality (37) has also been established in earlier works (Fazel et al., 2018; Fu et al., 2020).

Lemma 6 *Let K be an actor parameter such that $\rho(A - BK) < 1$, we have*

$$c_2\,\mathrm{Tr}(G_K G_K^\top) \leq J(K) - J(K^*) \leq c_3\,\mathrm{Tr}(G_K G_K^\top), \quad (37)$$

with positive constants $c_2 = \frac{\sigma_{min}(D_\epsilon)}{\|R\| + c_P\|B\|^2}$ and $c_3 = \frac{\|D_{K^*}\|}{\sigma_{min}(R)}$.

We recall that $\|\cdot\|$ denotes the operator norm of a matrix. We also recall that K^* is the optimal control parameter that is given by $K^* = (R + B^\top P^* B)^{-1}B^\top P^* A$ (see (2) for definition of P^*). Next lemma establishes the improvement of the actor update.

Lemma 7 (Improvement in the actor update) *Let the actor update be defined by* (25) *and Assumption 1 hold, then*

$$J(K_{t+1}) - J(K_t) \leq -\beta_t\frac{\sigma_{min}(D_\epsilon)}{c_3}(J(K_t) - J(K^*)) \\ - \beta_t\left[\sigma_{min}(D_\epsilon) - \beta_t c_D(\|R\| + c_P\|B\|^2)\right]\|\widehat{G}_{K_t}\|_F^2 + \beta_t c_D\|G_{K_t} - \widehat{G}_{K_t}\|_F^2$$

Remark 5 *This actor improvement lemma is a generalization of Lemma 15 in Fazel et al. (2018). Their lemma shows an improvement of policy gradient with accurate critic, while our lemma shows that there are extra terms when we have stochastic estimate of the critic.*

5. Numerical Examples

In this section, we present some numerical examples to validate our theoretical results. The code can be found at Zhou. We consider two examples: the first one has $d = 2$ and $k = 3$:

$$A = \begin{bmatrix}0.5 & 0\\ 0 & 0.5\end{bmatrix},\ B = \begin{bmatrix}0.2 & 0 & 0.1\\ 0 & 0.2 & 0.1\end{bmatrix},\ Q = \begin{bmatrix}1 & 0\\ 0 & 0.8\end{bmatrix},\ R = \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 0.5\end{bmatrix},\ D_\xi = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix},$$

and $\sigma = 1$. The other one has $d = 4$ and $k = 3$:

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

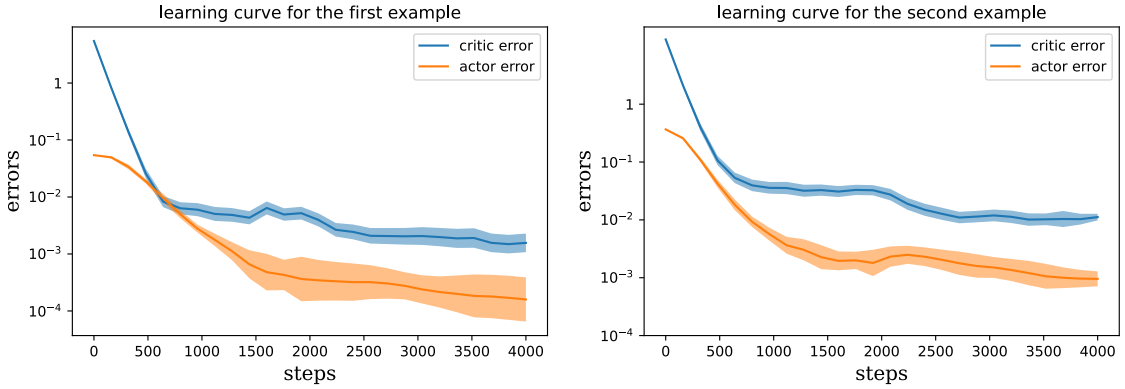

Figure 1: The error curves for the two examples with step size $\alpha_t = \beta_t = 0.001$. The errors are the average of 10 independent runs, with standard deviation plotted.

$$R = \begin{bmatrix} 1 & 0.1 & 0 \\ 0.1 & 1 & 0.1 \\ 0 & 0.1 & 1 \end{bmatrix}, \quad D_\xi = \begin{bmatrix} 1 & 0 & 0.1 & 0 \\ 0 & 1 & 0 & 0 \\ 0.1 & 0 & 1 & 0.1 \\ 0 & 0 & 0.1 & 1 \end{bmatrix},$$

and $\sigma = 1$. In all the tests, we set $N = N_0 = N_1 = 100$ for simplicity. We test for $T = 125, 250, 500, 1000, 2000, 4000$. In each example, we set the step sizes to be $\alpha_t = \beta_t = \frac{4}{T}$. In order to save time, we multiply the step sizes by 3 for the first $T/2$ steps.

Figure 1 shows the learning curves for the two example with step size $\alpha_t = \beta_t = 0.001$. The error is the average of 10 independent runs, and we also show the standard deviations. In the beginning, the error curves are nearly straight lines, which coincide with our one-step improvement analysis in the previous section. Then the errors become static because the algorithm has reached its capacity.

In order to obtain a convergence rate, we also test different step sizes, which is shown in Figure 2. In the tests, we keep $T\alpha_t = T\beta_t$ as a constant. The horizontal axis marks the number of steps T, ranging from 125 to 4000. We take a log_2 transform of T. The vertical axis is the final critic and actor errors (after a log_2 transform). A linear regression indicates that the slopes of the four error curves are all -1.0, which confirms our theoretical results in the previous section.

We also track the norm in Assumption 1. In the numerical tests, the maximum of $\rho(A - BK_t)$, $\|A - BK_t\|$, $\|E_t\|$, $\|K_t\|$, and $\|\theta_t\|_F$ for the first and second examples are 0.524, 0.529, 0.586, 0.329, 2.641 and 0.662, 0.662, 0.867, 0.498, 4.254 respectively. This further confirms that Assumption 1 is reasonable.

Acknowledgments

This work is supported in part by the National Science Foundation via grants DMS-2012286 and CCF-1934964 (Duke TRIPODS).

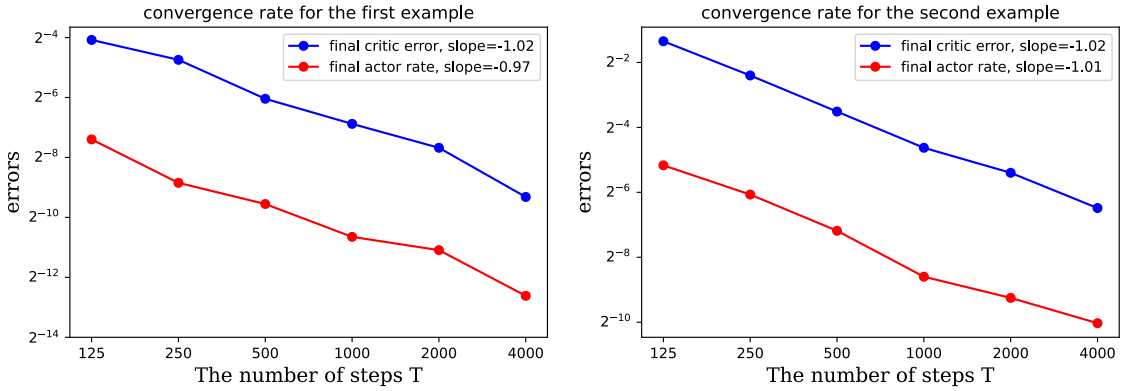

Figure 2: The convergence rate for the two examples with the numbers of steps ranging from $T = 125$ to $T = 4000$ and step size $\frac{4}{T}$. Each error is the average of 10 independent runs. The slope for the four error curves are all -1.0.

Appendix A. Proofs

Throughout the proof, we will frequently use two basic properties in linear algebra. So we state them here. The first one is that if X is a (symmetric and) positive semi-definite matrix and Y is of the same shape, then $\mathrm{Tr}(XY) \leq \mathrm{Tr}(X)\|Y\|$, where we recall that $\|\cdot\|$ is the operator norm of a matrix. The second property is a direct corollary of the first one: for any matrices X and Y of proper shapes, we have $\|XY\|_F \leq \|X\| \, \|Y\|_F$

A.1 Proofs for results in Section 2 and Section 3

Proof [Proof of Proposition 1] Since $\rho(A - BK) < 1$, we know from definition (8) that the expression for P_K in series is

$$P_K = \sum_{s=0}^{\infty} ((A - BK)^\top)^s (Q + K^\top R K)(A - BK)^s. \tag{38}$$

Give the state x_s, the conditional expectation of one-step cost is

$$\begin{aligned} \mathbb{E}[c(x_s, u_s)|x_s] &= x_s^\top Q x_s + \mathbb{E}_{\omega_s \sim N(0,Id)}[(-Kx_s + \sigma\omega_s)^\top R(-Kx_s + \sigma\omega_s)] \\ &= x_s^\top (Q + K^\top R K) x_s + \sigma^2 \, \mathrm{Tr}(R). \end{aligned} \tag{39}$$

So the total cost is

$$\begin{aligned}
J(K) &= \lim_{S\to\infty} \mathbb{E}_K \left[\frac{1}{S}\sum_{s=0}^{S-1} c(x_s,u_s)\right] = \lim_{S\to\infty} \mathbb{E}_K \left[\frac{1}{S}\sum_{s=0}^{S-1} \mathbb{E}[c(x_s,u_s)|x_s]\right] \\
&= \lim_{S\to\infty} \mathbb{E}_K \left[\frac{1}{S}\sum_{s=0}^{S-1} x_s^\top (Q+K^\top RK)x_s\right] + \sigma^2 \operatorname{Tr}(R) \\
&= \mathbb{E}_K[x^\top (Q+K^\top RK)x] + \sigma^2 \operatorname{Tr}(R) \\
&= \operatorname{Tr}\left[\mathbb{E}_K[xx^\top](Q+K^\top RK)\right] + \sigma^2 \operatorname{Tr}(R) = \operatorname{Tr}\left[D_K(Q+K^\top RK)\right] + \sigma^2 \operatorname{Tr}(R) \\
&= \operatorname{Tr}\left[D_K(P_K - (A-BK)^\top P_K(A-BK))\right] + \sigma^2 \operatorname{Tr}(R) \\
&= \operatorname{Tr}\left[(D_K - (A-BK)D_K(A-BK)^\top)P_K\right] + \sigma^2 \operatorname{Tr}(R) = \operatorname{Tr}[D_\epsilon P_K] + \sigma^2 \operatorname{Tr}(R).
\end{aligned}$$

So (9) holds. Next, we derive the expression for $\nabla_K J(K)$. We need a simple formula: if the shape of M is the same as the shape of K, then $\nabla_K \operatorname{Tr}(M^\top K) = \nabla_K \operatorname{Tr}(MK^\top) = M$. Since $J(K) = \operatorname{Tr}\left[D_K(Q+K^\top RK)\right] + \sigma^2 \operatorname{Tr}(R)$, we have

$$\nabla_K J(K) = 2RKD_K + \nabla_K \operatorname{Tr}[D_K Q_0]|_{Q_0 = Q+K^\top RK}. \tag{40}$$

We recall that

$$D_K = D_\epsilon + (A-BK)D_K(A-BK)^\top.$$

Therefore,

$$\begin{aligned}
&\nabla_K \operatorname{Tr}[D_K Q_0] = \nabla_K \operatorname{Tr}[(D_\epsilon + (A-BK)D_K(A-BK)^\top)Q_0] \\
&= -B^\top(Q_0+Q_0^\top)(A-BK)D_K + \nabla_K \operatorname{Tr}[D_K Q_1]|_{Q_1=(A-BK)^\top Q_0(A-BK)} \\
&= -2B^\top Q_0(A-BK)D_K + \nabla_K \operatorname{Tr}[D_K Q_1]|_{Q_1=(A-BK)^\top Q_0(A-BK)}
\end{aligned} \tag{41}$$

where we used $Q_0 = Q_0^\top$ in the last equality. Therefore, we can apply (41) recursively and obtain

$$\begin{aligned}
&\nabla_K \operatorname{Tr}[D_K Q_0]|_{Q_0=Q+K^\top RK} \\
&= -2B^\top(Q+K^\top RK)(A-BK)D_K + \nabla_K \operatorname{Tr}[D_K Q_1]|_{Q_1=(A-BK)^\top(Q+K^\top RK)(A-BK)} \\
&= -2B^\top(Q+K^\top RK)(A-BK)D_K - 2B^\top(A-BK)^\top(Q+K^\top RK)(A-BK)^2 D_K \\
&\quad + \nabla_K \operatorname{Tr}[D_K Q_2]|_{Q_2=((A-BK)^\top)^2(Q+K^\top RK)(A-BK)^2} \\
&= \cdots \\
&= -\sum_{s=0}^{\infty} 2B^\top((A-BK)^\top)^s(Q+K^\top RK)(A-BK)^{s+1}D_K \\
&= -2B^\top P_K(A-BK)D_K
\end{aligned} \tag{42}$$

where the assumption $\rho(A-BK)<1$ guarantees that the series converges and the remaining term vanishes. Substituting (42) into (40), we obtain

$$\nabla_K J(K) = 2RKD_K - 2B^\top P_K(A-BK)D_K = 2\left[(R+B^\top P_K B)K - B^\top P_K A\right]D_K.$$

■

Proof [Proof of Proposition 2] If we start with $x_0 = x$, since the state dynamic is

$$x_{s+1} = (A - BK)x_s + \epsilon_s$$

with $\epsilon_s \sim N(0, D_\epsilon)$, the state distribution is

$$x_s \sim N\left((A-BK)^s x,\ \sum_{i=0}^{s-1}(A-BK)^i D_\epsilon((A-BK)^\top)^i\right) =: N\left((A-BK)^s x,\ D_K^{(s)}\right).$$

Therefore, by definition, the value function is

$$\begin{aligned}
V_K(x) &= \sum_{s=0}^{\infty}\left\{\mathbb{E}_K\left[c(x_s,u_s) \mid x_0 = x\right] - J(K)\right\} \\
&= \sum_{s=0}^{\infty}\left\{\mathbb{E}_K\left[x_s^\top(Q + K^\top R K)x_s \mid x_0 = x\right] + \sigma^2 \operatorname{Tr}(R) - J(K)\right\} \\
&= \sum_{s=0}^{\infty}\left\{\operatorname{Tr}\left(\mathbb{E}_K\left[x_s x_s^\top \mid x_0 = x\right](Q + K^\top R K)\right) - \operatorname{Tr}[D_\epsilon P_K]\right\} \\
&= \sum_{s=0}^{\infty}\left\{\operatorname{Tr}\left[\left((A-BK)^s x x^\top((A-BK)^\top)^s + D_K^{(s)}\right)(Q + K^\top R K)\right] - \operatorname{Tr}[D_\epsilon P_K]\right\},
\end{aligned}$$

where the second equality is by (39), the third equality is by (9). Therefore,

$$\begin{aligned}
&V_K(x) \\
&= x^\top P_K x + \sum_{s=0}^{\infty}\left\{\operatorname{Tr}\left[\left(\sum_{i=0}^{s-1}(A-BK)^i D_\epsilon((A-BK)^\top)^i\right)(Q + K^\top R K)\right]\right. \\
&\qquad\qquad \left. - \operatorname{Tr}\left[D_\epsilon\left(\sum_{i=0}^{\infty}((A-BK)^\top)^i(Q + K^\top R K)(A-BK)^i\right)\right]\right\} \\
&= x^\top P_K x - \sum_{s=0}^{\infty}\operatorname{Tr}\left[\left(\sum_{i=s}^{\infty}(A-BK)^i D_\epsilon((A-BK)^\top)^i\right)(Q + K^\top R K)\right] \\
&= x^\top P_K x - \sum_{s=0}^{\infty}\sum_{j=0}^{\infty}\operatorname{Tr}\left[\left((A-BK)^s D_\epsilon((A-BK)^\top)^s\right)\right. \\
&\qquad\qquad \left.\left(((A-BK)^\top)^j(Q + K^\top R K)(A-BK)^j\right)\right] \\
&= x^\top P_K x - \sum_{s=0}^{\infty}\left\{\operatorname{Tr}\left[\left((A-BK)^s D_\epsilon((A-BK)^\top)^s\right)P_K\right]\right\} = x^\top P_K x - \operatorname{Tr}[D_K P_K],
\end{aligned}$$

where we have used the series expressions for P_K (38) and D_K (5). The assumption $\rho(A-BK) < 1$ guarantees that all the series above converge. Next, we compute the state-value function $Q_K(x,u)$.

Recall that $Q_K(x,u)$ is the expected extra cost if we start at $x_0 = x$, take a first action $u_0 = u$ and then follow the policy π_K. Therefore,

$$\begin{aligned}
&Q_K(x,u) = c(x,u) - J(K) + \mathbb{E}[V_K(x') \mid x,u] \\
&= x^\top Q x + u^\top R u - \operatorname{Tr}[D_\epsilon P_K] - \sigma^2 \operatorname{Tr}(R) + \mathbb{E}_{x' \sim N(Ax+Bu, D_\xi)}[x'^\top P_K x' - \operatorname{Tr}[D_K P_K]] \\
&= x^\top Q x + u^\top R u - \operatorname{Tr}[D_\epsilon P_K] - \sigma^2 \operatorname{Tr}(R) + \operatorname{Tr}\left[\mathbb{E}_{x' \sim N(Ax+Bu, D_\xi)}[x' x'^\top] P_K\right] - \operatorname{Tr}[D_K P_K] \\
&= x^\top Q x + u^\top R u - \operatorname{Tr}[D_\epsilon P_K + \sigma^2 R + D_K P_K] + \operatorname{Tr}\left[\left((Ax+Bu)(Ax+Bu)^\top + D_\xi\right) P_K\right] \\
&= x^\top Q x + u^\top R u - \operatorname{Tr}[(D_\epsilon - D_\xi) P_K + \sigma^2 R + D_K P_K] + (Ax+Bu)^\top P_K (Ax+Bu) \\
&= \begin{bmatrix} x^\top & u^\top \end{bmatrix} \begin{bmatrix} Q + A^\top P_K A & A^\top P_K B \\ B^\top P_K A & R + B^\top P_K B \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} - \sigma^2 \operatorname{Tr}(R + P_K B B^\top) - \operatorname{Tr}(D_K P_K).
\end{aligned}$$

■

Proof [Proof of Proposition 3] The distribution of policy is $\pi_K(u|x) \sim N(-Kx, \sigma^2 I_k)$, with probability density

$$\pi_K(u|x) = (2\pi\sigma^2)^{-k/2} \exp\left(-\frac{1}{2\sigma^2}|u + Kx|^2\right).$$

Therefore,

$$\log \pi_K(u|x) = -\frac{k}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}|u+Kx|^2$$

and

$$\nabla_K \log \pi_K(u|x) = -\frac{1}{\sigma^2}(u+Kx)x^\top.$$

Therefore, by the definition in (23), the Fisher information matrix at state x is

$$\begin{aligned}
F_x(K) &= \int_{\mathbb{R}^k} (2\pi\sigma^2)^{-k/2} \exp\left(-\frac{1}{2\sigma^2}|u+Kx|^2\right) \frac{1}{\sigma^4}[(u+Kx)x^\top] \otimes [(u+Kx)x^\top] \,\mathrm{d}u \\
&= \int_{\mathbb{R}^k} (2\pi\sigma^2)^{-k/2} \exp\left(-\frac{1}{2\sigma^2}|u|^2\right) \frac{1}{\sigma^4}[ux^\top] \otimes [ux^\top] \,\mathrm{d}u.
\end{aligned}$$

Recall that the stationary state distribution is $N(0, D_K)$. Hence, the Fisher information matrix is

$$\begin{aligned}
F(K) &= \int_{\mathbb{R}^d} (2\pi)^{-d/2} (\det(D_K))^{-1/2} \exp\left(-\frac{1}{2}x^\top D_K^{-1} x\right) F_x(K) \,\mathrm{d}x \\
&= \int_{\mathbb{R}^d} (2\pi)^{-d/2} (\det(D_K))^{-1/2} \exp\left(-\frac{1}{2}x^\top D_K^{-1} x\right) \\
&\quad \int_{\mathbb{R}^k} (2\pi\sigma^2)^{-k/2} \exp\left(-\frac{1}{2\sigma^2}|u|^2\right) \frac{1}{\sigma^4}[ux^\top] \otimes [ux^\top] \,\mathrm{d}u \,\mathrm{d}x
\end{aligned}$$

Note that we can compute the integration w.r.t. x and u separately with

$$\int_{\mathbb{R}^d} (2\pi)^{-d/2} (\det(D_K))^{-1/2} \exp\left(-\frac{1}{2}x^\top D_K^{-1} x\right) xx^\top \,\mathrm{d}x = D_K$$

and

$$\int_{\mathbb{R}^k} (2\pi\sigma^2)^{-k/2} \exp\left(-\frac{1}{2\sigma^2}|u|^2\right) uu^\top \, \mathrm{d}u = \sigma^2 I_k.$$

Therefore, by an elementwise analysis, we obtain

$$\sigma^2 F(K) \cdot G_K = G_K D_K.$$

Therefore, (24) holds. ■

A.2 Proofs for results in section 4

We first prove the lemmas and then the main theorem 1.

Proof [Proof of Lemma 1] Firstly

$$D_{K_t} = D_\epsilon + (A - BK_t) D_{K_t} (A - BK_t)^\top \geq D_\epsilon \geq \sigma_{min}(D_\epsilon).$$

D_{K_t} also has an expression in series:

$$D_{K_t} = \sum_{s=0}^{\infty} (A - BK_t)^s D_\epsilon ((A - BK_t)^\top)^s.$$

Since $\lim_{k\to\infty} \|(A-BK_t)^k\|^{\frac{1}{k}} = \rho(A-BK_t) \leq \rho < 1$ and $\|A - BK_t\| \leq c_A$, (with an argument similar to the proof in Lemma 2 below,) we have

$$D_{K_t} = \sum_{s=0}^{\infty} (A - BK_t)^s D_\epsilon ((A - BK_t)^\top)^s \lesssim \frac{1}{1-\rho^2} \|D_\epsilon\|$$

with the constant depending on c_A and d. Therefore, the first inequality in (32) holds. The constant c_D is proportional to $\frac{1}{1-\rho^2}\|D_\epsilon\|$ and also depends on c_A and d. The argument above also holds for K^*, so the inequality also holds with K_t replaced by K^*. For P_{K_t}, we also have an expression in series:

$$P_{K_t} = \sum_{s=0}^{\infty} ((A - BK_t)^\top)^s (Q + B^\top R B)(A - BK_t)^s.$$

So the argument to prove the second inequality of (32) is the same. Finally, since Σ_{K_t} has expression (13) with $\|D_{K_t}\| \leq c_D$ and $\|K_t\| \leq c_K$, $\|\Sigma_{K_t}\|$ has a bound $c_\Sigma = (1 + c_K)^2 c_D + \sigma^2$ automatically. ■

A.2.1 PROOFS FOR CRITIC

Here we prove the results for the critic.

Proof [Proof of Lemma 2] In order to show $\nabla^2 L_K(\theta) = \mathbb{E}_K [\psi(x,u) \otimes \psi(x,u)] \geq \mu_\sigma$, we only need to show that for any $M \in \mathbb{R}^{(d+k)\times(d+k)}$, we have

$$\mathbb{E}_K \left[(\mathrm{Tr}[M\psi(x,u)])^2\right] \geq \mu_\sigma \|M\|_F^2.$$

Since $\psi(x,u)$ is symmetric, we have $\mathrm{Tr}[M\psi(x,u)] = \mathrm{Tr}[M^\top\psi(x,u)] = \mathrm{Tr}[\frac{1}{2}(M+M^\top)\psi(x,u)]$. We also have $2\|\frac{1}{2}(M+M^\top)\|_F^2 \geq \|M\|_F^2$. Therefore, we only need to show

$$\mathbb{E}_K\left[(\mathrm{Tr}[M\psi(x,u)])^2\right] \geq 2\mu_\sigma\|M\|_F^2 \tag{43}$$

for all symmetric matrix M. Recall that

$$z_{s+1} = Ez_s + \widetilde{\epsilon}_s.$$

Since

$$\psi(z) = \mathbb{E}_K[(Ez+\widetilde{\epsilon})(Ez+\widetilde{\epsilon})^\top] - zz^\top = Ezz^\top E^\top + \Sigma_\epsilon - zz^\top,$$

we have

$$\mathrm{Tr}[M\psi(x,u)] = \mathrm{Tr}[MEzz^\top E^\top + M\Sigma_\epsilon - Mzz^\top] = z^\top(E^\top ME - M)z + \mathrm{Tr}[M\Sigma_\epsilon].$$

Recall that $z \sim N(0,\Sigma_K)$ under the stationary distribution where Σ_K is defined in (13). By definition, for any $x \in \mathbb{R}^d$, $u \in \mathbb{R}^k$, and $\gamma \neq 0$, we have

$$\begin{aligned}\begin{bmatrix}x^\top & u^\top\end{bmatrix}\Sigma_K\begin{bmatrix}x\\u\end{bmatrix} = (\gamma x - \frac{1}{\gamma}K^\top u)^\top D_K(\gamma x - \frac{1}{\gamma}K^\top u)\\ + (1-\gamma^2)x^\top D_K x + u^\top[\sigma^2 I_k - (\frac{1}{\gamma^2}-1)KD_KK^\top]u.\end{aligned} \tag{44}$$

Therefore, we can smartly choose a $\gamma \in (0,1)$ s.t. $(1-\gamma^2)D_K \geq \mu_\Sigma$ and $\sigma^2 I_k - (\frac{1}{\gamma^2}-1)KD_KK^\top \geq \mu_\Sigma$ for some positive constant $\mu_\Sigma \in \mathbb{R}$. Therefore, $\Sigma_K \geq \mu_\Sigma$. Using the same method, we can also show that $\Sigma_\epsilon \geq \mu_\Sigma$. This μ_Σ depends on σ, $\sigma_{min}(D_K)$ ($\sigma_{min}(D_\epsilon)$ for Σ_ϵ) and $\|K\|$. Since $\sigma_{min}(D_K) \geq \sigma_{min}(D_\epsilon) = \mathcal{O}(1)$, μ_Σ is of order $\mathcal{O}(1)$ as long as we have an upper bound for $\|K\|$. We can also find that $\mu_\Sigma = \mathcal{O}(\sigma^2)$ when σ is small. Next, we start to compute (43).

$$\begin{aligned}&\mathbb{E}_K\left[(\mathrm{Tr}[M\psi(x,u)])^2\right]\\ &= \mathbb{E}_K\left[\left(z^\top(E^\top ME - M)z + \mathrm{Tr}[M\Sigma_\epsilon]\right)\left(z^\top(E^\top ME - M)z + \mathrm{Tr}[M\Sigma_\epsilon]\right)\right]\\ &= \mathbb{E}_K\left[z^\top(E^\top ME - M)zz^\top(E^\top ME - M)z + 2z^\top(E^\top ME - M)z\,\mathrm{Tr}[M\Sigma_\epsilon] + \mathrm{Tr}[M\Sigma_\epsilon]^2\right].\end{aligned} \tag{45}$$

We will compute each term respectively. We recall the stationary distribution is $z \sim N(0,\Sigma_K)$. If we define $w = \Sigma_K^{-\frac{1}{2}}z$, then $w \sim N(0, I_{d+k})$. Denote $(m_{ij}) = \widetilde{M} = \Sigma_K^{\frac{1}{2}}(E^\top ME - M)\Sigma_K^{\frac{1}{2}}$, then

$\widetilde{M}$ is symmetric and

$$\begin{aligned}
&\mathbb{E}_K\left[z^\top(E^\top ME-M)zz^\top(E^\top ME-M)z\right]\\
&=\mathbb{E}_{w\sim N(0,I_{d+k})}\left[w^\top\Sigma_K^{\frac{1}{2}}(E^\top ME-M)\Sigma_K^{\frac{1}{2}}ww^\top\Sigma_K^{\frac{1}{2}}(E^\top ME-M)\Sigma_K^{\frac{1}{2}}w\right]\\
&=\mathbb{E}_{w\sim N(0,I_{d+k})}\left[w^\top\widetilde{M}ww^\top\widetilde{M}w\right]\\
&=\int_{\mathbb{R}^{d+k}}(2\pi)^{-\frac{d+k}{2}}w^\top\widetilde{M}ww^\top\widetilde{M}w\exp\left(-\frac{|w|^2}{2}\right)\mathrm{d}w\\
&=3\sum_{i=1}^{d+k}m_{ii}^2+\sum_{i\neq j}m_{ii}m_{jj}+2\sum_{i\neq j}m_{ij}^2=2\,\mathrm{Tr}[\widetilde{M}^2]+\mathrm{Tr}[\widetilde{M}]^2\\
&=2\,\mathrm{Tr}\left[\Sigma_K(E^\top ME-M)\Sigma_K(E^\top ME-M)\right]+\mathrm{Tr}\left[\Sigma_K(E^\top ME-M)\right]^2.
\end{aligned}\tag{46}$$

Also,

$$\mathbb{E}_K\left[z^\top(E^\top ME-M)z\right]=\mathbb{E}_K\left[\mathrm{Tr}(zz^\top(E^\top ME-M))\right]=\mathrm{Tr}[\Sigma_K(E^\top ME-M))].\tag{47}$$

Recall that $\Sigma_K=\Sigma_\epsilon+E\Sigma_KE^\top$, so

$$\mathrm{Tr}[\Sigma_K(E^\top ME-M))]=-\mathrm{Tr}[M(\Sigma_K-E\Sigma_KE^\top)]=-\mathrm{Tr}[M\Sigma_\epsilon]\tag{48}$$

Therefore, substituting (46), (47) and (48) into (45), we obtain

$$\begin{aligned}
&\mathbb{E}_K\left[(\mathrm{Tr}[M\psi(x,u)])^2\right]\\
&=2\,\mathrm{Tr}\left[\Sigma_K(E^\top ME-M)\Sigma_K(E^\top ME-M)\right]+\mathrm{Tr}\left[M\Sigma_\epsilon\right]^2-2\,\mathrm{Tr}\left[M\Sigma_\epsilon\right]^2+\mathrm{Tr}\left[M\Sigma_\epsilon\right]^2\\
&=2\,\mathrm{Tr}\left[\Sigma_K(E^\top ME-M)\Sigma_K(E^\top ME-M)\right]\\
&\geq 2\mu_\Sigma\,\mathrm{Tr}\left[(E^\top ME-M)\Sigma_K(E^\top ME-M)\right]\\
&\geq 2\mu_\Sigma^2\left\|E^\top ME-M\right\|_F^2
\end{aligned}\tag{49}$$

for all symmetric matrix M. Next, we want to show $\|M\|_F\lesssim\|E^\top ME-M\|_F$. Since the Frobenius norm is equivalent to the operator norm (with the constant depending on the dimension), we only need to show $\|M\|\lesssim\|E^\top ME-M\|$. Note that this step makes μ_σ depend polynomially on $d+k$. We define an operator $\mathcal{T}_E:\mathbb{R}^{(d+k)\times(d+k)}\to\mathbb{R}^{(d+k)\times(d+k)}$ such that

$$\mathcal{T}_E(X)=\sum_{s=0}^{\infty}(E^\top)^sXE^s.$$

Since $1>\rho\geq\rho(E)=\lim_{s\to\infty}\|E^s\|^{\frac{1}{s}}$, the norm of the operator should satisfy

$$\|\mathcal{T}_E\|=sup_{X\neq 0}\frac{\|\mathcal{T}_E(X)\|}{\|X\|}\leq\frac{c}{1-\rho^2}$$

where c depends on $\|E\|$ and $d+k$. Notice that

$$\mathcal{T}_E(M - E^\top ME) = \sum_{s=0}^{\infty} (E^\top)^s (M - E^\top ME) E^s = M,$$

we conclude that

$$\|M\| = \|\mathcal{T}_E(M - E^\top ME)\| \leq \|\mathcal{T}_E\| \|M - E^\top ME\| \leq \frac{c}{1-\rho^2} \|M - E^\top ME\|.$$

So, $\|M\|_F \lesssim \|E^\top ME - M\|_F$. Therefore, by (49), $\nabla^2 L_K(\theta) = \mathbb{E}_K [\psi(x,u) \otimes \psi(x,u)] \geq \mu_\sigma$ holds with μ_σ proportional to $\sigma^4/(1-\rho^2)$ and depending on $\|E\|$ and $d+k$. Moreover, μ_σ grows polynomially as $d+k$ becomes large. ■

Proof [Proof of Lemma 3] Similar to (19), we define

$$\nabla L_t(\theta_t) = \mathbb{E}_{K_t}[f(x,u)],$$

where f depends on both θ_t and K_t. We denote $\mathbb{E}_{N_0}[f(x,u)]$ the expectation of the same function under the distribution of (x_{N_0}, u_{N_0}), which starts at $x_0 = 0$ and follows the policy π_{K_t}. We prove (34) first. We recall that the feature matrix $\phi(x,u)$ defined in (14) is quadratic in (x,u). So, $\psi(x,u) = \mathbb{E}[\phi(x',u')|x,u] - \phi(x,u)$ also grows at most quadratically in (x,u) since (x',u') are normally distributed. Therefore, $f(x,u)$, defined in (19) grows at most quartically in (x,u). By assumption 1, $\|\theta_t\|_F \leq c_\theta = \mathcal{O}(1)$ and $\|K_t\| \leq c_K = \mathcal{O}(1)$, so the coefficients for this quadratic growth are of order $\mathcal{O}(1)$. A similar argument tells us that $\widehat{f}(x_{N_0}^{(i)}, u_{N_0}^{(i)})$ defined in (20) grows at most quartically in $\{(x^{(i,j)}, u^{(i,j)})\}_{j=1}^{N_1}$ and $(x_{N_0}^{(i)}, u_{N_0}^{(i)})$, with $\mathcal{O}(1)$ coefficients. Note that $\{(x^{(i,j)}, u^{(i,j)})\}_{j=1}^{N_1}$ and $(x_{N_0}^{(i)}, u_{N_0}^{(i)})$ are normally distributed with 0 mean and $\mathcal{O}(1)$ covariance matrix. Therefore,

$$\mathbb{E}\left[\left\| \widehat{\nabla L_t}(\theta_t) \right\|_F^2 \,\middle|\, \mathcal{G}_t \right] = \mathbb{E}\left[\left\| \frac{1}{N} \sum_{i=1}^{N} \widehat{f}(x_{N_0}^{(i)}, u_{N_0}^{(i)}) \right\|_F^2 \,\middle|\, \mathcal{G}_t \right] = \mathcal{O}(1)$$

So (34) holds with $c_L = \mathcal{O}(1)$. We also see that $c_L = \text{poly}(d+k)$ as the dimensions increase. We will show (33) next. By definition,

$$\left\| \mathbb{E}\left[\widehat{\nabla L_t}(\theta_t) - \nabla L_t(\theta_t) \,\middle|\, \mathcal{G}_t \right] \right\|_F = \|\mathbb{E}_{N_0}[f(x,u)] - \mathbb{E}_{K_t}[f(x,u)]\|_F. \tag{50}$$

Here, we remind the reader that the expectation on the left in (50) is taken w.r.t. the training filtration $\mathcal{G}_t$ while those on the right are taken w.r.t. the state and action distributions.

We remark that existing results (Arnold and Avez, 1968) bound (50) directly. However, it can be computed directly, so we give an elementary proof. Recall that the state trajectory is given by

$$x_{s+1} = (A - BK_t)x_s + \epsilon_s$$

with $x_0 = 0$ where $\epsilon_s \sim N(0, D_\epsilon)$. Therefore, the distribution of x_{N_0} is

$$x_{N_0} \sim N\left(0, \sum_{s=0}^{N_0-1} (A - BK_t)^s D_\epsilon ((A - BK_t)^\top)^s \right) =: N\left(0, D_{K_t}^{(N_0)}\right)$$

and the stationary distribution of x_s is

$$x_\infty \sim N\left(0, \sum_{s=0}^{\infty}(A-BK_t)^s D_\epsilon((A-BK_t)^\top)^s\right) = N\left(0, D_{K_t}\right).$$

Since $\rho(A-BK_t) \le \rho < 1$, $D_\epsilon > 0$, and $N_0 = \mathcal{O}(\log \frac{1}{\delta})$, we have $D_{K_t} > \sigma_{min}(D_\epsilon)$, $D_{K_t}^{(N_0)} > \sigma_{min}(D_\epsilon)$, $D_{K_t} - D_{K_t}^{(N_0)} \ge 0$ and $\|D_{K_t} - D_{K_t}^{(N_0)}\|_F \lesssim \delta$. Since $u_s \sim N(-K_t x_s, \sigma^2 I_k)$, we have the joint distribution for $z_{N_0} = (x_{N_0}^\top, u_{N_0}^\top)^\top$

$$z_{N_0} \sim N\left(0, \begin{bmatrix} D_{K_t}^{(N_0)} & -D_{K_t}^{(N_0)} K_t^\top \\ -K_t D_{K_t}^{(N_0)} & K_t D_{K_t}^{(N_0)} K_t^\top + \sigma^2 I_k \end{bmatrix}\right) =: N\left(0, \Sigma_{K_t}^{(N_0)}\right)$$

and the joint stationary distribution

$$z \sim N\left(0, \begin{bmatrix} D_{K_t} & -D_{K_t} K_t^\top \\ -K_t D_{K_t} & K_t D_{K_t} K_t^\top + \sigma^2 I_k \end{bmatrix}\right) =: N\left(0, \Sigma_{K_t}\right)$$

Since $\|D_{K_t} - D_{K_t}^{(N_0)}\|_F \lesssim \delta$ and $\|K_t\| \le c_K$, we have $\|\Sigma_{K_t} - \Sigma_{K_t}^{(N_0)}\|_F \le c_6 \delta$. Here the positive constant $c_6 = \mathcal{O}(1)$ decrease geometrically as N_0 increases algebraically. Furthermore, using the same argument when we prove $\Sigma_K \ge \mu_\Sigma$ in Lemma 2, we can find a positive constant $\mu_\Sigma = \mathcal{O}(1)$ such that $\Sigma_{K_t} \ge \mu_\Sigma$ and $\Sigma_{K_t}^{(N_0)} \ge \mu_\Sigma$. Therefore

$$\begin{aligned}
&\|\mathbb{E}_{N_0}[f(x,u)] - \mathbb{E}_{K_t}[f(x,u)]\|_F \\
=& \left\| \int_{\mathbb{R}^{d+k}} f(z)(2\pi)^{-\frac{d+k}{2}} \left[\det(\Sigma_{K_t}^{(N_0)})^{-\frac{1}{2}} \exp\left(-\frac{1}{2} z^\top (\Sigma_{K_t}^{(N_0)})^{-1} z\right) \right.\right. \\
&\qquad \left.\left. - \det(\Sigma_{K_t})^{-\frac{1}{2}} \exp\left(-\frac{1}{2} z^\top (\Sigma_{K_t})^{-1} z\right) \right] \mathrm{d}z \right\|_F \\
\le& \int_{\mathbb{R}^{d+k}} c(1+|z|^4) \left| \det(\Sigma_{K_t}^{(N_0)})^{-\frac{1}{2}} \exp\left(-\frac{1}{2} z^\top (\Sigma_{K_t}^{(N_0)})^{-1} z\right) \right. \\
&\qquad \left. - \det(\Sigma_{K_t})^{-\frac{1}{2}} \exp\left(-\frac{1}{2} z^\top (\Sigma_{K_t})^{-1} z\right) \right| \mathrm{d}z \\
\le& \int_{\mathbb{R}^{d+k}} c(1+|z|^4) \left[\det(\Sigma_{K_t}^{(N_0)})^{-\frac{1}{2}} - \det(\Sigma_{K_t})^{-\frac{1}{2}} \right] \exp\left(-\frac{1}{2} z^\top (\Sigma_{K_t}^{(N_0)})^{-1} z\right) \mathrm{d}z \\
&+ \int_{\mathbb{R}^{d+k}} c(1+|z|^4) \det(\Sigma_{K_t})^{-\frac{1}{2}} \left[\exp\left(-\frac{1}{2} z^\top (\Sigma_{K_t})^{-1} z\right) - \exp\left(-\frac{1}{2} z^\top (\Sigma_{K_t}^{(N_0)})^{-1} z\right) \right] \mathrm{d}z
\end{aligned} \tag{51}$$

There is no absolute value at the end of (51) because each term is non-negative. Next, we will bound the two integrals respectively. For the first one, we have

$$\begin{aligned}
&\det(\Sigma_{K_t}^{(N_0)})^{-\frac{1}{2}} - \det(\Sigma_{K_t})^{-\frac{1}{2}} \\
=& \frac{\det(\Sigma_{K_t}) - \det(\Sigma_{K_t}^{(N_0)})}{\sqrt{\det(\Sigma_{K_t}^{(N_0)}) \det(\Sigma_{K_t})} \left(\sqrt{\det(\Sigma_{K_t})} + \sqrt{\det(\Sigma_{K_t}^{(N_0)})} \right)} \\
=& \mathcal{O}(1) \left(\det(\Sigma_{K_t}) - \det(\Sigma_{K_t}^{(N_0)}) \right).
\end{aligned}$$

Next, we will show $\det(\Sigma_{K_t}) - \det(\Sigma_{K_t}^{(N_0)}) = \mathcal{O}(\delta)$. We can find a unitary matrix U such that $U^\top \Sigma_{K_t}^{(N_0)} U$ is a diagonal matrix,

$$\left\| U^\top \Sigma_{K_t} U - U^\top \Sigma_{K_t}^{(N_0)} U \right\|_F = \left\| \Sigma_{K_t} - \Sigma_{K_t}^{(N_0)} \right\|_F \leq c_6 \delta,$$

and

$$\det(\Sigma_{K_t}) - \det(\Sigma_{K_t}^{(N_0)}) = \det(U \Sigma_{K_t} U^\top) - \det(U \Sigma_{K_t}^{(N_0)} U^\top).$$

If we assume that the diagonal element of $U \Sigma_{K_t} U^\top$ to be $a_1, \cdots, a_{d+k}$ and

$$U \Sigma_{K_t}^{(N_0)} U^\top = \text{diag}(b_1, \cdots, b_{d+k}).$$

Then $a_i \geq b_i$ and $a_i - b_i = \mathcal{O}(\delta)$. Therefore

$$0 \leq \det(U \Sigma_{K_t} U^\top) - \det(U \Sigma_{K_t}^{(N_0)} U^\top) \leq \prod_{i=1}^{d+k} a_i - \prod_{i=1}^{d+k} b_i = \mathcal{O}(\delta).$$

Therefore, $\det(\Sigma_{K_t}) - \det(\Sigma_{K_t}^{(N_0)}) = \mathcal{O}(\delta)$ and hence

$$\det(\Sigma_{K_t}^{(N_0)})^{-\frac{1}{2}} - \det(\Sigma_{K_t})^{-\frac{1}{2}} \leq c\delta$$

with positive constant c being as small as we want (through increasing N_0). Therefore, the first integral in (51) satisfies

$$\begin{aligned} &\int_{\mathbb{R}^{d+k}} c(1+|z|^4) \left[\det(\Sigma_{K_t}^{(N_0)})^{-\frac{1}{2}} - \det(\Sigma_{K_t})^{-\frac{1}{2}} \right] \exp\left(-\frac{1}{2} z^\top (\Sigma_{K_t}^{(N_0)})^{-1} z \right) \mathrm{d}z \\ &\leq c\delta \int_{\mathbb{R}^{d+k}} (1+|z|^4) \exp\left(-\frac{1}{2} z^\top (\Sigma_{K_t}^{(N_0)})^{-1} z \right) \mathrm{d}z = c\delta \mathcal{O}(1) \leq \frac{1}{2}\delta. \end{aligned} \tag{52}$$

Here, again, the constant c may differ according to the context. A more detailed computation shows that

$$\det(\Sigma_{K_t}^{(N_0)})^{-\frac{1}{2}} - \det(\Sigma_{K_t})^{-\frac{1}{2}} \leq \det(\Sigma_{K_t}^{(N_0)})^{-\frac{1}{2}} \, \text{poly}(d+k) \, c_6 \delta.$$

Therefore, N_0 should scale with $\log(d+k)$ as the dimensions increase. Next, we bound the second integration in (51). Using the inequality $1 - e^{-x} \leq x$, we have

$$\begin{aligned} &\exp\left(-\frac{1}{2} z^\top (\Sigma_{K_t})^{-1} z \right) - \exp\left(-\frac{1}{2} z^\top (\Sigma_{K_t}^{(N_0)})^{-1} z \right) \\ &= \exp\left(-\frac{1}{2} z^\top (\Sigma_{K_t})^{-1} z \right) \left[1 - \exp\left(-\frac{1}{2} z^\top \left((\Sigma_{K_t}^{(N_0)})^{-1} - (\Sigma_{K_t})^{-1} \right) z \right) \right] \\ &\leq \frac{1}{2} z^\top \left((\Sigma_{K_t}^{(N_0)})^{-1} - (\Sigma_{K_t})^{-1} \right) z \exp\left(-\frac{1}{2} z^\top (\Sigma_{K_t})^{-1} z \right) \\ &= \frac{1}{2} \exp\left(-\frac{1}{2} z^\top (\Sigma_{K_t})^{-1} z \right) \text{Tr}\left[\left((\Sigma_{K_t}^{(N_0)})^{-1} - (\Sigma_{K_t})^{-1} \right) z z^\top \right] \\ &= \frac{1}{2} \exp\left(-\frac{1}{2} z^\top (\Sigma_{K_t})^{-1} z \right) \text{Tr}\left[(\Sigma_{K_t}^{(N_0)})^{-1} \left(\Sigma_{K_t} - \Sigma_{K_t}^{(N_0)} \right) (\Sigma_{K_t})^{-1} z z^\top \right] \\ &\leq \frac{1}{2} \exp\left(-\frac{1}{2} z^\top (\Sigma_{K_t})^{-1} z \right) \| (\Sigma_{K_t}^{(N_0)})^{-1} \left(\Sigma_{K_t} - \Sigma_{K_t}^{(N_0)} \right) (\Sigma_{K_t})^{-1} \| \, \text{Tr}[z z^\top] \\ &\leq \frac{1}{2} \exp\left(-\frac{1}{2} z^\top (\Sigma_{K_t})^{-1} z \right) \frac{1}{\mu_\Sigma^2} c_6 \delta |z|^2. \end{aligned}$$

Therefore, the second integration in (51) satisfies

$$\begin{aligned}&\int_{\mathbb{R}^{d+k}} c(1+|z|^4)\det(\Sigma_{K_t})^{-\frac{1}{2}}\left[\exp\left(-\frac{1}{2}z^\top(\Sigma_{K_t})^{-1}z\right)-\exp\left(-\frac{1}{2}z^\top(\Sigma_{K_t}^{(N_0)})^{-1}z\right)\right]\mathrm{d}z\\ &\leq \delta\int_{\mathbb{R}^{d+k}} c(|z|^2+|z|^6)\det(\Sigma_{K_t})^{-\frac{1}{2}}\exp\left(-\frac{1}{2}z^\top(\Sigma_{K_t})^{-1}z\right)\mathrm{d}z=\delta c\mathcal{O}(1)\leq\frac{1}{2}\delta.\end{aligned} \tag{53}$$

Plugging (52) and (53) into (51), we obtain

$$\|\mathbb{E}_{N_0}[f(x,u)]-\mathbb{E}_{K_t}[f(x,u)]\|_F\leq\delta.$$

■

Proof [Proof of Lemma 4] By definition

$$\theta_K-\theta_{K'}=\begin{bmatrix}A^\top(P_K-P_{K'})A & A^\top(P_K-P_{K'})B\\ B^\top(P_K-P_{K'})A & B^\top(P_K-P_{K'})B\end{bmatrix}=\begin{bmatrix}A^\top\\ B^\top\end{bmatrix}\left[P_K-P_{K'}\right]\begin{bmatrix}A & B\end{bmatrix}$$

Therefore,

$$\begin{aligned}&\|\theta_K-\theta_{K'}\|_F^2=\mathrm{Tr}[(\theta_K-\theta_{K'})^\top(\theta_K-\theta_{K'})]\\ &=\mathrm{Tr}\left([(AA^\top+BB^\top)(P_K-P_{K'})]^2\right)\leq(\|A\|^2+\|B\|^2)^2\|P_K-P_{K'}\|_F^2\end{aligned} \tag{54}$$

Therefore, our goal is to bound $\|P_K-P_{K'}\|_F$ by $\|K-K'\|_F$. By definition in (8),

$$\begin{aligned}&P_K-P_{K'}\\ &=K^\top RK-K'^\top RK'+(A-BK)^\top P_K(A-BK)-(A-BK')^\top P_{K'}(A-BK')\\ &=K^\top RK-K^\top RK'+K^\top RK'-K'^\top RK'\\ &\quad+(A-BK)^\top P_K(A-BK)-(A-BK)^\top P_K(A-BK')\\ &\quad+(A-BK)^\top P_K(A-BK')-(A-BK)^\top P_{K'}(A-BK')\\ &\quad+(A-BK)^\top P_{K'}(A-BK')-(A-BK')^\top P_{K'}(A-BK')\\ &=K^\top R(K-K')+(K-K')^\top RK'-(A-BK)^\top P_KB(K-K')\\ &\quad+(A-BK)^\top(P_K-P_{K'})(A-BK')-(K-K')^\top B^\top P_{K'}(A-BK')\end{aligned}$$

Therefore,

$$\begin{aligned}&P_K-P_{K'}-(A-BK)^\top(P_K-P_{K'})(A-BK')\\ &=K^\top R(K-K')+(K-K')^\top RK'\\ &\quad-(A-BK)^\top P_KB(K-K')-(K-K')^\top B^\top P_{K'}(A-BK')\end{aligned} \tag{55}$$

Next, we want to take $\|\cdot\|_F$ on both sides of (55). For the left hand side, since $\rho(A-BK),\rho(A-BK')\leq\rho<1$ and $\|A-BK\|,\|A-BK'\|\leq c_A$, we can repeat the last part in the proof of Lemma 2 and prove that

$$\|P_K-P_{K'}\|_F\leq c\|(P_K-P_{K'})-(A-BK)^\top(P_K-P_{K'})(A-BK')\|_F \tag{56}$$

where c is proportional to $1/(1-\rho^2)$ and also depends on c_A and d. For the right hand side of (55), since $\|P_K\| \le c_P$, $\|P_{K'}\| \le c_P$, $\|K\| \le c_K$ and $\|K'\| \le c_K$,

$$\begin{aligned} &\|K^\top R(K-K') + (K-K')^\top RK' \\ &\quad - (A-BK)^\top P_K B(K-K') - (K-K')^\top B^\top P_{K'}(A-BK')\|_F \\ &\le 2(c_K\|R\| + c_P c_A\|B\|)\,\|K-K'\|_F. \end{aligned} \tag{57}$$

Plugging (56) and (57) into (55), we obtain

$$\|P_K - P_{K'}\|_F \le 2c(c_K\|R\| + c_P c_A\|B\|)\,\|K-K'\|_F. \tag{58}$$

Finally, combining (54) and (58), we obtain

$$\|\theta_K - \theta_{K'}\|_F \le c_1\|K-K'\|_F \tag{59}$$

with $c_1 = 2c(c_K\|R\| + c_P c_A\|B\|)\,(\|A\|^2 + \|B\|^2)$. This c_1 grows polynomially as the dimensions increase. ■

Proof [Proof of Lemma 5] Note that

$$\begin{aligned} &\|\theta_{t+1} - \theta_{K_{t+1}}\|_F^2 = \|\theta_t - \alpha_t\widehat{\nabla L_t}(\theta_t) - \theta_{K_t} + \theta_{K_t} - \theta_{K_{t+1}}\|_F^2 \\ &= \|\theta_t - \theta_{K_t}\|_F^2 - 2\alpha_t \operatorname{Tr}\left[(\theta_t - \theta_{K_t})^\top \widehat{\nabla L_t}(\theta_t)\right] \\ &\quad + \alpha_t^2\|\widehat{\nabla L_t}(\theta_t)\|_F^2 + \|\theta_{K_t} - \theta_{K_{t+1}}\|_F^2 + 2\operatorname{Tr}\left[(\theta_{K_t} - \theta_{K_{t+1}})^\top(\theta_t - \theta_{K_t} - \alpha_t\widehat{\nabla L_t}(\theta_t))\right] \\ &= \|\theta_t - \theta_{K_t}\|_F^2 - 2\alpha_t \operatorname{Tr}\left[(\theta_t - \theta_{K_t})^\top \nabla L_t(\theta_t)\right] + 2\alpha_t \operatorname{Tr}\left[(\theta_t - \theta_{K_t})^\top(\nabla L_t(\theta_t) - \widehat{\nabla L_t}(\theta_t))\right] \\ &\quad + \alpha_t^2\|\widehat{\nabla L_t}(\theta_t)\|_F^2 + \|\theta_{K_t} - \theta_{K_{t+1}}\|_F^2 + 2\operatorname{Tr}\left[(\theta_{K_t} - \theta_{K_{t+1}})^\top(\theta_t - \theta_{K_t} - \alpha_t\widehat{\nabla L_t}(\theta_t))\right] \\ &\le (1 - 2\alpha_t\mu_\sigma)\|\theta_t - \theta_{K_t}\|_F^2 + 2\alpha_t \operatorname{Tr}\left[(\theta_t - \theta_{K_t})^\top(\nabla L_t(\theta_t) - \widehat{\nabla L_t}(\theta_t))\right] + \alpha_t^2\|\widehat{\nabla L_t}(\theta_t)\|_F^2 \\ &\quad + \|\theta_{K_t} - \theta_{K_{t+1}}\|_F^2 + 2\operatorname{Tr}\left[(\theta_{K_t} - \theta_{K_{t+1}})^\top(\theta_t - \theta_{K_t})\right] - 2\alpha_t \operatorname{Tr}\left[(\theta_{K_t} - \theta_{K_{t+1}})^\top\widehat{\nabla L_t}(\theta_t)\right] \\ &\le (1 - \frac{5}{3}\alpha_t\mu_\sigma)\|\theta_t - \theta_{K_t}\|_F^2 + 2\alpha_t \operatorname{Tr}\left[(\theta_t - \theta_{K_t})^\top(\nabla L_t(\theta_t) - \widehat{\nabla L_t}(\theta_t))\right] + 2\alpha_t^2\|\widehat{\nabla L_t}(\theta_t)\|_F^2 \\ &\quad + (\frac{3}{\alpha_t\mu_\sigma} + 2)\|\theta_{K_t} - \theta_{K_{t+1}}\|_F^2 \end{aligned} \tag{60}$$

The first inequality is because $L_t(\theta)$ is $\mu_\sigma-$ strongly convex and hence

$$\operatorname{Tr}\left[(\theta_t - \theta_{K_t})^\top \nabla L_t(\theta_t)\right] = \operatorname{Tr}\left[(\theta_t - \theta_{K_t})^\top(\nabla L_t(\theta_t) - \nabla L_t(\theta_{K_t}))\right] \ge \mu_\sigma\|\theta_t - \theta_{K_t}\|_F^2.$$

The second inequality in (60) is a simple application of Cauchy-Schwartz inequality. Taking expectation w.r.t. $\mathcal{G}_t$ in (60), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\|\theta_{t+1}-\theta_{K_{t+1}}\|_F^2 \mid \mathcal{G}_t\right] \\
&\leq (1-\frac{5}{3}\alpha_t\mu_\sigma)\|\theta_t-\theta_{K_t}\|_F^2 + 2\alpha_t \operatorname{Tr}\left[(\theta_t-\theta_{K_t})^\top \mathbb{E}\left[\nabla L_t(\theta_t) - \widehat{\nabla L_t}(\theta_t) \;\middle|\; \mathcal{G}_t\right]\right] \\
&\quad + 2\alpha_t^2 \mathbb{E}\left[\left\|\widehat{\nabla L_t}(\theta_t)\right\|_F^2 \;\middle|\; \mathcal{G}_t\right] + (\frac{3}{\alpha_t\mu_\sigma}+2)\|\theta_{K_t}-\theta_{K_{t+1}}\|_F^2 \\
&\leq (1-\frac{4}{3}\alpha_t\mu_\sigma)\|\theta_t-\theta_{K_t}\|_F^2 + \frac{3\alpha_t}{\mu_\sigma}\left\|\mathbb{E}\left[\nabla L_t(\theta_t) - \widehat{\nabla L_t}(\theta_t) \;\middle|\; \mathcal{G}_t\right]\right\|_F^2 \\
&\quad + 2\alpha_t^2 \mathbb{E}\left[\left\|\widehat{\nabla L_t}(\theta_t)\right\|_F^2 \;\middle|\; \mathcal{G}_t\right] + (\frac{3}{\alpha_t\mu_\sigma}+2)\|\theta_{K_t}-\theta_{K_{t+1}}\|_F^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\|\theta_{t+1}-\theta_{K_{t+1}}\|_F^2 \mid \mathcal{G}_t\right] - \|\theta_t-\theta_{K_t}\|_F^2 \\
&\leq -\frac{4}{3}\alpha_t\mu_\sigma\|\theta_t-\theta_{K_t}\|_F^2 + \frac{3\alpha_t}{\mu_\sigma}\left\|\mathbb{E}\left[\nabla L_t(\theta_t) - \widehat{\nabla L_t}(\theta_t) \;\middle|\; \mathcal{G}_t\right]\right\|_F^2 \\
&\quad + 2\alpha_t^2 \mathbb{E}\left[\left\|\widehat{\nabla L_t}(\theta_t)\right\|_F^2 \;\middle|\; \mathcal{G}_t\right] + (\frac{3}{\alpha_t\mu_\sigma}+2)\|\theta_{K_t}-\theta_{K_{t+1}}\|_F^2.
\end{aligned}$$

Combining with (33), (34), and the definition of α_t, we obtain (36):

$$\begin{aligned}
&\mathbb{E}\left[\|\theta_{t+1}-\theta_{K_{t+1}}\|_F^2 \mid \mathcal{G}_t\right] - \|\theta_t-\theta_{K_t}\|_F^2 \\
&\leq -\frac{4}{3}\alpha_t\mu_\sigma\|\theta_t-\theta_{K_t}\|_F^2 + \frac{1}{4}\frac{\sigma_{min}(D_\epsilon)}{c_3}\beta_t\varepsilon + \big(\frac{3}{\alpha_t\mu_\sigma}+2\big)\|\theta_{K_t}-\theta_{K_{t+1}}\|_F^2.
\end{aligned}$$

■

A.2.2 PROOFS FOR THE ACTOR

Next, we prove the results for the actor.

Proof [Proof of Lemma 6] We prove the upper bound first. According to (9),

$$J(K)-J(K^*) = \operatorname{Tr}((P_K-P_{K^*})D_\epsilon) = \mathbb{E}_{x\sim N(0,D_\epsilon)}[x^\top(P_K-P_{K^*})x] \tag{61}$$

where we recall that $P_K = (Q+K^\top RK)+(A-BK)^\top P_K(A-BK)$ and P_{K^*} satisfies a similar equation. So, P_{K^*} also has the following expression in series

$$P_{K^*} = \sum_{s=0}^{\infty}[(A-BK^*)^s]^\top(Q+K^{*\top}RK^*)(A-BK^*)^s.$$

Therefore, if we define a sequence $\{y_s\}_{s=0}^\infty$ with $y_0=x$ and $y_{s+1}=(A-BK^*)y_s$, then

$$x^\top P_{K^*}x = \sum_{s=0}^{\infty} x^\top[(A-BK^*)^s]^\top(Q+K^{*\top}R^*K)(A-BK^*)^s x = \sum_{s=0}^{\infty} y_s^\top(Q+K^{*\top}RK^*)y_s.$$

Combining with

$$x^\top P_K x = \sum_{s=0}^{\infty} \left(y_s^\top P_K y_s - y_{s+1}^\top P_K y_{s+1} \right) = \sum_{s=0}^{\infty} y_s^\top (P_K - (A - BK^*)^\top P_K (A - BK^*)) y_s$$

and (61), we obtain

$$\begin{aligned} & J(K) - J(K^*) \\ &= \mathbb{E}_{D_\epsilon, K^*} \left[\sum_{s=0}^{\infty} y_s^\top \left(-Q - K^{*\top} R K^* + P_K - (A - BK^*)^\top P_K (A - BK^*) \right) y_s \right] \\ &= \mathrm{Tr} \left[\mathbb{E}_{D_\epsilon, K^*} \left[\sum_{s=0}^{\infty} y_s y_s^\top \right] \cdot \left(-Q - K^{*\top} R K^* + P_K - (A - BK^*)^\top P_K (A - BK^*) \right) \right] \end{aligned} \tag{62}$$

where $\mathbb{E}_{D_\epsilon, K^*}$ denotes the expectation with $y_0 \sim N(0, D_\epsilon)$ and $y_{s+1} = (A - BK^*) y_s$. Next, we analyze the two terms in (62) respectively. The first term is easy, recall that D_{K^*} is the solution of

$$D_{K^*} = D_\epsilon + (A - BK^*) D_{K^*} (A - BK^*)^\top$$

so that

$$D_{K^*} = \sum_{s=0}^{\infty} (A - BK^*)^s D_\epsilon [(A - BK^*)^\top]^s.$$

Therefore,

$$\mathbb{E}_{D_\epsilon, K^*} \left[\sum_{s=0}^{\infty} y_s y_s^\top \right] = \mathbb{E}_{x \sim N(0, D_\epsilon)} \left[\sum_{s=0}^{\infty} (A - BK^*)^s x x^\top [(A - BK^*)^\top]^s \right] = D_{K^*}. \tag{63}$$

Next, we consider the second term in (62). By direct computation,

$$\begin{aligned} & - Q - K^{*\top} R K^* + P_K - (A - BK^*)^\top P_K (A - BK^*) \\ &= -Q - (K^* - K + K)^\top R (K^* - K + K) + P_K \\ & \quad - (A - BK + BK - BK^*)^\top P_K (A - BK + BK - BK^*) \\ &= (K - K^*)^\top (RK - B^\top P_K (A - BK)) + (RK - B^\top P_K (A - BK))^\top (K - K^*) \\ & \quad - (K - K^*)^\top (R + B^\top P_K B)(K - K^*) \\ &= (K - K^*)^\top G_K + G_K^\top (K - K^*) - (K - K^*)^\top (R + B^\top P_K B)(K - K^*) \\ &= G_K^\top (R + B^\top P_K B)^{-1} G_K \\ & \quad - (K - K^* - (R + B^\top P_K B)^{-1} G_K)^\top (R + B^\top P_K B)(K - K^* - (R + B^\top P_K B)^{-1} G_K) \\ &\leq G_K^\top (R + B^\top P_K B)^{-1} G_K \end{aligned} \tag{64}$$

where we have used the equation (8) for P_K in the second equality and the definition of G_K (22) in the third equality. The $\leq$ above means the difference of the two matrix is positive semi-definite. Plugging (63) and (64) into (62), we obtain

$$J(K) - J(K^*) \leq \mathrm{Tr}(D_{K^*} G_K^\top (R + B^\top P_K B)^{-1} G_K) \leq \|D_{K^*}\| / \sigma_{min}(R) \, \mathrm{Tr}(G_K G_K^\top).$$

This finishes the proof of the upper bound. Next, we prove the lower bound. Note that the argument above does not rely on the optimality of K^*. Therefore, we can obtain a general formula (that is useful in the proof later):

$$\begin{aligned}&J(K) - J(K')\\ &= \operatorname{Tr}\left[D_{K'}\left((K-K')^\top G_K + G_K^\top (K-K') - (K-K')^\top (R+B^\top P_K B)(K-K')\right)\right]. \quad (65)\end{aligned}$$

Specifically, we can set $K' = K - (R+B^\top P_K B)^{-1} G_K$ (i.e., let (64) hold with equality), then by the optimality of K^* and (65), we obtain

$$\begin{aligned}&J(K) - J(K^*) \geq J(K) - J(K') = \operatorname{Tr}(D_{K'} G_K^\top (R+B^\top P_K B)^{-1} G_K)\\ &\geq \sigma_{min}(D_\epsilon)\, \|R+B^\top P_K B\|^{-1} \operatorname{Tr}(G_K G_K^\top) \geq \frac{\sigma_{min}(D_\epsilon)}{\|R\| + c_P \|B\|^2} \operatorname{Tr}(G_K G_K^\top)\end{aligned}$$

■

Proof [Proof of Lemma 7] By (65),

$$\begin{aligned}&J(K_t) - J(K_{t+1})\\ &= \operatorname{Tr}\left[D_{K_{t+1}}\left((K_t - K_{t+1})^\top G_{K_t} + G_{K_t}^\top (K_t - K_{t+1})\right.\right.\\ &\quad \left.\left. -(K_t - K_{t+1})^\top (R+B^\top P_{K_t} B)(K_t - K_{t+1})\right)\right]\\ &= \operatorname{Tr}\left[D_{K_{t+1}}\left(\beta_t \widehat{G}_{K_t}^\top G_{K_t} + \beta_t G_{K_t}^\top \widehat{G}_{K_t} - \beta_t^2 \widehat{G}_{K_t}^\top (R+B^\top P_{K_t} B)\widehat{G}_{K_t}\right)\right]\end{aligned}$$

Therefore,

$$\begin{aligned}&J(K_{t+1}) - J(K_t)\\ &= -\beta_t \operatorname{Tr}\left[D_{K_t}\left(\widehat{G}_{K_t}^\top G_{K_t} + G_{K_t}^\top \widehat{G}_{K_t} - \beta_t \widehat{G}_{K_t}^\top (R+B^\top P_{K_t} B)\widehat{G}_{K_t}\right)\right]\\ &= -\beta_t \operatorname{Tr}\left[D_{K_t}\left(G_{K_t}^\top G_{K_t} + \widehat{G}_{K_t}^\top \widehat{G}_{K_t} - (G_{K_t} - \widehat{G}_{K_t})^\top (G_{K_t} - \widehat{G}_{K_t}) - \beta_t \widehat{G}_{K_t}^\top (R+B^\top P_{K_t} B)\widehat{G}_{K_t}\right)\right]\end{aligned}$$

Recall that we proved

$$\sigma_{min}(D_\epsilon) I_d \leq D_{K_t} \leq c_D I_d \text{ and } P_{K_t} \leq c_P$$

in Lemma 1. Therefore,

$$\operatorname{Tr}\left[D_{K_t} G_{K_t}^\top G_{K_t}\right] \geq \sigma_{min}(D_\epsilon) \|G_{K_t}\|_F^2,$$

$$\operatorname{Tr}\left[D_{K_t} \widehat{G}_{K_t}^\top \widehat{G}_{K_t}\right] \geq \sigma_{min}(D_\epsilon) \|\widehat{G}_{K_t}\|_F^2,$$

$$\operatorname{Tr}\left[D_{K_t} \widehat{G}_{K_t}^\top (R+B^\top P_{K_t} B)\widehat{G}_{K_t}\right] \leq c_D(\|R\| + c_P\|B\|^2)\|\widehat{G}_{K_t}\|_F^2,$$

and

$$\operatorname{Tr}\left[D_{K_t}(G_{K_t} - \widehat{G}_{K_t})^\top (G_{K_t} - \widehat{G}_{K_t})\right] \leq c_D \|G_{K_t} - \widehat{G}_{K_t}\|_F^2.$$

Therefore,

$$J(K_{t+1}) - J(K_t) \leq -\beta_t \sigma_{min}(D_\epsilon)(\|G_{K_t}\|_F^2 + \|\widehat{G}_{K_t}\|_F^2) + \beta_t c_D \|G_{K_t} - \widehat{G}_{K_t}\|_F^2 + \beta_t^2 c_D(\|R\| + c_P\|B\|^2)\|\widehat{G}_{K_t}\|_F^2$$

Finally, by Lemma 6, we can conclude that

$$J(K_{t+1}) - J(K_t) \leq -\beta_t \frac{\sigma_{min}(D_\epsilon)}{c_3}(J(K_t) - J(K^*)) - \beta_t \left[\sigma_{min}(D_\epsilon) - \beta_t c_D(\|R\| + c_P\|B\|^2)\right] \|\widehat{G}_{K_t}\|_F^2 + \beta_t c_D \|G_{K_t} - \widehat{G}_{K_t}\|_F^2$$

■

A.2.3 PROOFS FOR THE MAIN THEOREM

Finally we can prove our main theorem.

Proof [Proof of Theorem 1] By lemma 3, (33) and (34) hold for all $t \leq T$. We define a Lyapunov function

$$\mathcal{L}_t = \mathcal{L}(\theta_t, K_t) = \|\theta_t - \theta_{K_t}\|_F^2 + J(K_t) - J(K^*).$$

Firstly, $\mathcal{L}_0 = \mathcal{O}(1)$ because

$$\|\theta_0 - \theta_{K_0}\|_F^2 = \|\theta_{K_0}\|_F^2 = \left\| \begin{bmatrix} Q + A^\top P_{K_0} A & A^\top P_{K_0} B \\ B^\top P_{K_0} A & R + B^\top P_{K_0} B \end{bmatrix} \right\|_F^2 = \mathcal{O}(1)$$

(note that $P_{K_0} = Q + A^\top P_{K_0} A$ implies $\|P_{K_0}\|_F = \mathcal{O}(1)$) and

$$J(K_0) - J(K^*) \leq J(K_0) = \operatorname{Tr}(D_\epsilon P_{K_0}) + \sigma^2 \operatorname{Tr}(R) \leq c_P \operatorname{Tr}[D_\epsilon] + \sigma^2 \operatorname{Tr}(R) = \mathcal{O}(1).$$

Next, we want to show a decrease rate of the Lyapunov function. According to Lemma 5 and Lemma 7,

$$\begin{aligned} &\mathbb{E}\left[\mathcal{L}_{t+1} \mid \mathcal{G}_t\right] - \mathcal{L}_t \\ \leq& -\frac{4}{3}\alpha_t \mu_\sigma \|\theta_t - \theta_{K_t}\|_F^2 + \frac{1}{4}\frac{\sigma_{min}(D_\epsilon)}{c_3}\beta_t \varepsilon + (\frac{3}{\alpha_t \mu_\sigma} + 2)\|\theta_{K_t} - \theta_{K_{t+1}}\|_F^2 \\ &- \beta_t \frac{\sigma_{min}(D_\epsilon)}{c_3}(J(K_t) - J(K^*)) - \beta_t \left[\sigma_{min}(D_\epsilon) - \beta_t c_D(\|R\| + c_P\|B\|^2)\right] \|\widehat{G}_{K_t}\|_F^2 \\ &+ \beta_t c_D \|G_{K_t} - \widehat{G}_{K_t}\|_F^2. \end{aligned} \tag{66}$$

Fortunately, we can use the negative term in the actor estimate to bound the positive term in the critic estimate and use the negative term in the critic estimate to bound the positive term in the actor estimate. Specifically, by Lemma 4,

$$\|\theta_{K_t} - \theta_{K_{t+1}}\|_F^2 \leq c_1^2 \|K_t - K_{t+1}\|_F^2 = c_1^2 \beta_t^2 \|\widehat{G}_{K_t}\|_F^2.$$

So, by the second inequality in (28)

$$\beta_t \left[\sigma_{min}(D_\epsilon) - \beta_t c_D(\|R\| + c_P\|B\|^2)\right] \|\widehat{G}_{K_t}\|_F^2 \geq (\frac{3}{\alpha_t \mu_\sigma} + 2)\|\theta_{K_t} - \theta_{K_{t+1}}\|_F^2. \tag{67}$$

In addition,

$$\|G_{K_t} - \widehat{G}_{K_t}\|_F^2 = \|(\theta_{K_t}^{22} - \theta_t^{22})K_t - (\theta_{K_t}^{21} - \theta_t^{21})\|_F^2 \leq c_K^2 \|\theta_t - \theta_{K_t}\|_F^2.$$

So, by the third inequality in (28)

$$\frac{1}{3}\alpha_t \mu_\sigma \|\theta_t - \theta_{K_t}\|_F^2 \geq \beta_t c_D \|G_{K_t} - \widehat{G}_{K_t}\|_F^2. \tag{68}$$

Substituting (67) and (68) into (66), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{L}_{t+1} \mid \mathcal{G}_t\right] - \mathcal{L}_t \\
&\leq -\alpha_t \mu_\sigma \|\theta_t - \theta_{K_t}\|_F^2 + \frac{1}{4}\frac{\sigma_{min}(D_\epsilon)}{c_3}\beta_t \varepsilon - \beta_t \frac{\sigma_{min}(D_\epsilon)}{c_3}(J(K_t) - J(K^*)).
\end{aligned}$$

Taking expectation, we obtain

$$\mathbb{E}[\mathcal{L}_{t+1} - \mathcal{L}_t] \leq -\mathbb{E}\left[\alpha_t \mu_\sigma \|\theta_t - \theta_{K_t}\|_F^2 + \beta_t \frac{\sigma_{min}(D_\epsilon)}{c_3}(J(K_t) - J(K^*))\right] + \frac{1}{4}\frac{\sigma_{min}(D_\epsilon)}{c_3}\beta_t \varepsilon. \tag{69}$$

Next, we consider three cases. The first case is when $\mathbb{E}[\|\theta_t - \theta_{K_t}\|_F^2] \geq \frac{1}{2}\varepsilon$. In this case, by (69) and the first inequality of (28),

$$\mathbb{E}[\mathcal{L}_{t+1} - \mathcal{L}_t] \leq -\mathbb{E}\left[\frac{1}{3}\alpha_t \mu_\sigma \|\theta_t - \theta_{K_t}\|_F^2 + \beta_t \frac{\sigma_{min}(D_\epsilon)}{c_3}(J(K_t) - J(K^*))\right].$$

The second case is when $\mathbb{E}[J(K_t) - J(K^*)] \geq \frac{1}{2}\varepsilon$. In this case

$$\mathbb{E}[\mathcal{L}_{t+1} - \mathcal{L}_t] \leq -\mathbb{E}\left[\alpha_t \mu_\sigma \|\theta_t - \theta_{K_t}\|_F^2 + \frac{1}{2}\beta_t \frac{\sigma_{min}(D_\epsilon)}{c_3}(J(K_t) - J(K^*))\right].$$

In both the first and the second cases, we have

$$\mathbb{E}[\mathcal{L}_{t+1} - \mathcal{L}_t] \leq -\mathbb{E}\left[\frac{1}{3}\alpha_t \mu_\sigma \|\theta_t - \theta_{K_t}\|_F^2 + \frac{1}{2}\beta_t \frac{\sigma_{min}(D_\epsilon)}{c_3}(J(K_t) - J(K^*))\right].$$

Note that $\frac{1}{2}\beta_t \frac{\sigma_{min}(D_\epsilon)}{c_3} \leq \frac{1}{3}\alpha_t \mu_\sigma$, we obtain a contraction rate for the Lyaponov function in both cases:

$$\mathbb{E}[\mathcal{L}_{t+1} - \mathcal{L}_t] \leq -\frac{1}{2}\beta_t \frac{\sigma_{min}(D_\epsilon)}{c_3}\mathbb{E}[\mathcal{L}_t] =: -\beta_t c_4 \mathbb{E}[\mathcal{L}_t]$$

where we remind the reader that $L(\theta_{K^*}, K^*) = 0$. Let us rewrite it into a contraction form

$$\mathbb{E}[\mathcal{L}_{t+1}] \leq (1 - \beta_t c_4)\mathbb{E}[\mathcal{L}_t]. \tag{70}$$

Next, we consider the third case, when both $\mathbb{E}[\|\theta_t - \theta_{K_t}\|_F^2] < \frac{1}{2}\varepsilon$ and $\mathbb{E}[J(K_t) - J(K^*)] < \frac{1}{2}\varepsilon$. In this case we have $\mathbb{E}[\mathcal{L}_t] < \varepsilon$. Therefore, by (69), we obtain

$$\begin{aligned}
&\mathbb{E}[\mathcal{L}_{t+1}] \\
&\leq (1 - \alpha_t \mu_\sigma)\mathbb{E}\left[\|\theta_t - \theta_{K_t}\|_F^2\right] + \frac{1}{4}\frac{\sigma_{min}(D_\epsilon)}{c_3}\beta_t \varepsilon + \left(1 - \beta_t \frac{\sigma_{min}(D_\epsilon)}{c_3}\right)\mathbb{E}\left[(J(K_t) - J(K^*))\right] \\
&< \frac{1}{2}\varepsilon + \frac{1}{2}\varepsilon\left(\frac{1}{2}\frac{\sigma_{min}(D_\epsilon)}{c_3}\beta_t + 1 - \beta_t \frac{\sigma_{min}(D_\epsilon)}{c_3}\right) < \varepsilon.
\end{aligned}$$

Therefore, we have shown that under (33) and (34), the Lyapunov function is decreasing at rate (70) as long as $\mathbb{E}[\|\theta_t - \theta_{K_t}\|_F^2] \geq \frac{1}{2}\varepsilon$ or $\mathbb{E}[J(K_t) - J(K^*)] \geq \frac{1}{2}\varepsilon$, or else, the Lyapunov function will keep being smaller than ε. Since $(1-\beta_t c_4)^T \mathcal{L}_0 < \varepsilon$ (recall that β_t is constant in t), we have $\mathbb{E}[\mathcal{L}_T] \leq \varepsilon$. Since $\mathbb{E}[\mathcal{L}_T]$ is the sum of two non-negative numbers, both of them are less than ε. ■

References

Brian DO Anderson and John B Moore. *Optimal control: linear quadratic methods*. Courier Corporation, 2007.

Vladimir Igorevich Arnold and André Avez. *Ergodic problems of classical mechanics*, volume 9. Benjamin, 1968.

Jonathan F Bard. *Practical bilevel optimization: algorithms and applications*, volume 30. Springer Science & Business Media, 2013.

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

Shalabh Bhatnagar, Richard S Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer, 2012.

Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1):33–57, 1996.

Tianyi Chen, Yuejiao Sun, Quan Xiao, and Wotao Yin. A single-timescale method for stochastic bilevel optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2466–2488. PMLR, 2022.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4): 633–679, 2020.

OS Ebrahim, MF Salem, PK Jain, and MA Badr. Application of linear quadratic regulator theory to the stator field-oriented control of induction motors. *IET Electric Power Applications*, 4(8): 637–646, 2010.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476. PMLR, 2018.

Zuyue Fu, Zhuoran Yang, and Zhaoran Wang. Single-timescale actor-critic provably finds globally optimal policy. *arXiv preprint arXiv:2008.00483*, 2020.

Walter R Gilks, Sylvia Richardson, and David Spiegelhalter. *Markov chain Monte Carlo in practice*. CRC press, 1995.

Aamir Hashim. Optimal speed control for direct current motors using linear quadratic regulator. *Journal of Engineering and Computer Science (JECS)*, 14(2):48–56, 2019.

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

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

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

Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. In *NeurIPS*, 2020.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. *Advances in Neural Information Processing Systems*, 32, 2019.

Hesameddin Mohammadi, Armin Zare, Mahdi Soltanolkotabi, and Mihailo R Jovanovic. Convergence and sample complexity of gradient methods for the model-free linear quadratic regulator problem. *IEEE Transactions on Automatic Control*, 2021.

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

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Ankur Sinha, Pekka Malo, and Kalyanmoy Deb. A review on bilevel optimization: from classical to evolutionary approaches and applications. *IEEE Transactions on Evolutionary Computation*, 22(2):276–295, 2017.

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

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

Marco A Wiering. Multi-agent reinforcement learning for traffic light control. In *Machine Learning: Proceedings of the Seventeenth International Conference (ICML'2000)*, pages 1151–1158, 2000.

Yue Frank Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite-time analysis of two time-scale actor-critic methods. *Advances in Neural Information Processing Systems*, 33:17617–17628, 2020.

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

Jiongmin Yong and Xun Yu Zhou. *Stochastic controls: Hamiltonian systems and HJB equations*, volume 43. Springer Science & Business Media, 1999.

Sihan Zeng, Thinh T Doan, and Justin Romberg. A two-time-scale stochastic optimization framework with applications in control and reinforcement learning. *arXiv preprint arXiv:2109.14756*, 2021.

Mo Zhou. Single time-scale actor-critic method to solve the linear quadratic regulator with convergence proof. `https://github.com/MoZhou1995/ActorCriticLQR.git`.