Proceedings of the 36th Chinese Control Conference
July 26-28, 2017, Dalian, China

Off-policy Reinforcement Learning for Robust Control of Discrete-time Uncertain Linear Systems

Yongliang Yang[1], Zhishan Guo[2], Donald Wunsch[3], Yixin Yin[1]

1. School of Automatic and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, P. R. China.
E-mail: y.yang.2016@ieee.org,yyx@ies.ustb.edu.cn

2. Department of Computer Science, Missouri University of Science and Technology, Rolla, MO 65401, USA.
E-mail: guozh@mst.edu

3. Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla, MO 65401, USA.
E-mail: wunsch@ieee.org

Abstract: In this paper, an off-policy reinforcement learning method is developed for the robust stabilizing controller design of discrete-time uncertain linear systems. The proposed robust control design consists of two steps. First, the robust control problem is transformed to an optimal control problem. Second, the off-policy RL method is used to design the optimal control policy which guarantees the robust stability of the original system with uncertainty. The condition for the equivalence between the robust control problem and the optimal control problem is discussed. The off-policy does not require any knowledge of the system knowledge and efficiently utilize the data collected from on-line to improve the performance of approximate optimal control policy in each iteration successively. Finally, a simulation example is carried out to verify the effectiveness of the presented algorithm for the robust control problem of discrete-time linear system with uncertainty.

Key Words: system uncertainty, robust control, optimal control, off-policy trinforcement learning, model-free

1 Introduction

The models of real world physical systems are usually coupled with model uncertainty, which is challenging to the feedback control design. Robust control is designed to deal with the uncertain parameters and structures within a certain bound in order to achieve the guaranteed performance. In the early time, robust control design was based on frequency domain analysis [1]. Since then, time domain based approaches were also developed to investigate the robust stabilization problem for both linear and nonlinear systems [2, 3]. The adaptive control techniques have been applied successfully to deal with multi-agent systems [4], time delay systems [5–7], and so on.

Another kind of robust control approach is proposed based on the optimal control design method. In [8], the robust control problem is transformed to an optimal control problem with a modified system dynamics and a performance function. The robust control problem is equivalent to the optimal control problem in the sense that the unique optimal control of the transformed optimal control problem is able to robustly stabilize the original system with uncertainty. In this way, the robust control problem reduces to the solution of the Hamilton-Jacobi-Bellman (HJB) equation for a general nonlinear system and the Riccati equation for the linear system. This idea has been successfully applied to guaranteed cost regulation problem in [9], guaranteed cost tracking problem in [10], robust control of uncertain constrained systems in [11], robust optimal control design in [12, 13]. It is necessary to develop an efficient method for the transformed optimal control problem which is able to equivalently tackle the robust control problem. This is the motivation for this paper.

An efficient method, referred to as approximate/adaptive dynamic programming (ADP) or reinforcement learning (RL), was proposed in [14]. The core idea of ADP, as the name indicates, is to find the solution that satisfies the HJB equation. For linear systems, the solution of Riccati equation was approximated successively by solving a sequence of Lyapunov equations in [15]. This idea was extended to nonlinear continuous-time systems in [16]. Since then, ADP developed from off-line [17] to on-line [18, 19], from model-based [20] to unknown systems [21]. A novel RL approach, called off-policy RL, was presented to approximate the optimal control policy in an on-line manner by a novel policy iteration algorithm for the continuous-time linear systems without any knowledge of the system dynamics [22]. Then, the off-policy RL method was extended to H_∞ control problem in [23, 24], tracking control problem in [25] and the output synchronization problem of multi-agent system in [26]. To the authors' knowledge, the off-policy RL method has not been applied to the robust control problem for discrete-time uncertain linear systems yet. The main contribution of this paper is to utilize the off-policy RL approach to solve the robust control problem of discrete-time uncertain linear systems without requiring any knowledge of the systems dynamics.

The remainder of this paper is organized as follows. Section 2 describes the robust control problem of discrete-time linear system with uncertainty. In Section 3, the robust control problem is transformed to the optimal control problem of a modified system. The condition for the equivalence that the optimal control policy of the modified system can robustly stabilize the origin uncertain system is also given in Section 3. The model-free off-policy RL method to solve the optimal control problem is described in Section 4. In Section 5, a simulation is conducted to demonstrate the validity of the proposed approach. Finally, concluding remarks

This work was supported in part by the Mary K. Finley Missouri Endowment, the Missouri S&T Intelligent Systems Center, the National Science Foundation, the National Natural Science Foundation of China (NSFC Grant No. 61333002) and the China Scholarship Council (CSC No. 201406460057).

and future works are presented in Section 5.

2 Problem Statement

Consider a class of discrete-time linear systems with uncertainty

$$x_{k+1} = [A + \Delta(p)] x_k + Bu_k, \tag{1}$$

where $x_k \in \mathbb{R}^n$ is the system state, $u_k \in \mathbb{R}^{m_1}$ is the control input, $A + \Delta \in \mathbb{R}^{n\times n}$ represents the drift dynamics and $B \in \mathbb{R}^{n\times m_1}$ is the input dynamics. $\Delta(p)$ with $p \in \Omega_p$ represents the bounded perturbation of the nominal system

$$x_{k+1} = Ax_k + Bu_k. \tag{2}$$

Without loss of generality, it is assumed that (A, B_1) is stabilizable, and $x = 0$ is an equilibrium state of system (1).

The system uncertainty can be classified into two types, the matched and mismatched uncertainty [8]. The class of matched uncertainty satisfies

$$\Delta(p) = B\phi(p). \tag{3}$$

That is, the matched uncertainty $\Delta(p)$ in (3) can only describe the class of uncertainty in the space spanned by the columns of the input matrix B_1. Therefore, another type of uncertainty, unmatched uncertainty, is considered

$$\Delta(p) = BB^{\dagger}\Delta(p) + \left(I - BB^{\dagger}\right)\Delta(p) \tag{4}$$

where $B_1^{\dagger}$ is the pseudo-inverse of B_1. In (4), the uncertainty $\Delta(p)$ is composed of the matched component $B_1B_1^{\dagger}\Delta(p)$ and the unmatched component $\left(I - B_1B_1^{\dagger}\right)\Delta(p)$. Assume that the uncertainty $\Delta(p)$ is upper bounded by the following inequality

$$\varepsilon^{-1}\Delta^T(p)\Delta(p) \le F,\ \forall p \in \Omega_p, \tag{5}$$

where $\varepsilon \succ 0$ is a design parameter to be determined.

The robust control problem of system (1) is formulated as the following.

(Robust Control Problem) To find a state feedback control law $u_k = Kx_k$ such that the close-loop system

$$\begin{aligned} x_{k+1} &= (A + BK)x_k + \Delta(p)x_k \\ &= A_c x_k + \Delta(p)x_k \end{aligned} \tag{6}$$

is asymptotically stable for $\forall p \in \Omega_p$.

In order to solve the robust control problem, the feedback gain can be obtained by an optimal control design method described as following.

(Optimal Control Problem) Consider the modified nominal system

$$x_{k+1} = Ax_k + Bu_k + Dv_k \tag{7}$$

with the performance described as

$$\begin{aligned} J(x_k, u_k) &= \frac{1}{2}\sum_{j=k}^{\infty}\left(x_j^T Q x_j + x_j^T F x_j + \beta^2 x_j^T x_j\right. \\ &+ \left. u_j^T R_1 u_j + v_j^T R_2 v_j\right) \end{aligned} \tag{8}$$

where $D = \alpha\left(I - B_1B_1^{\dagger}\right) \in \mathbb{R}^{n\times r}$ and r is the rank of B. $V(x_k)$ is also referred to as the value function [27]. The scalars $\alpha > 0$, $\beta > 0$ and the matrices $Q \succ 0$, $R_1 \succ 0$ and $R_2 \succ 0$ are parameters to be determined. For simplicity, denote the utility function as

$$\begin{aligned} r(x_k, u_k, v_k) &= x_k^T Q x_k + x_k^T F x_k + \beta^2 x_k^T x_k \\ &+ u_k^T R_1 u_k + v_k^T R_2 v_k. \end{aligned} \tag{9}$$

The optimal control problem of system (7) with respect to the performance (8) is to find the optimal state feedback control law $u_k^* = K^* x_k$ and $v_k^* = L^* x_k$, such that the value function $V(x_k)$ in (8) is minimized.

Remark 1. *As shown later in Section 3, under some specific conditions, K^* is able to stabilize the uncertain system (1), i.e. the optimal solution of the corresponding optimal control problem can stabilize the uncertain system. Note that the control input v_k only appears in the modified nominal system (7). Therefore, the feedback gain L^* does not affect the system (1) directly. However, L^* helps to design K^* to stabilize the system (1) indirectly.*

In the next section, it will be proved that the optimal feedback gain of system (7) with respect to the performance (8), K^*, is a stabilizing feedback gain for the uncertain system (1). Then the model-free ADP technique can be further employed to the corresponding optimal control problem.

3 Optimal Control Design Approach for the Robust Control Problem

In this section, the optimal control design based approach for the robust control problem is considered. First, the robust control problem is transformed to an optimal control problem for the system (7) with respect to the performance (8). The optimality condition for the optimal control and the expression of the optimal feedback gain K^* and L^* are derived. Then the conditions that guarantee the asymptotic stability of the closed-loop system (6) when the feedback gain K^* is given.

First, the definition of admissible control is required.

Definition 1. (Admissible Control) *For the modified nominal system (7), the control mappings $u(x)$ and $v(x)$ are said to be admissible with respect to performance (8) if 1) $u(x)$ and $v(x)$ are continuous; 2) $u(0) = v(0) = 0$; 3) $u(x)$ and $v(x)$ stabilize the modified nominal system (7); 4) the value function $V(x_k)$ is finite for $\forall x_k$.*

In the optimal control problem of system (7) with respect to (8), the objective is to find the optimal control u_k^* such that

$$V^*(x_k) = \min_{u_k} J(x_k, u_k) \tag{10}$$

The Bellman equation for the value function $V(x_k)$ in (8) is

$$V(x_k) = V(x_{k+1}) + r(x_k, u_k, v_k). \tag{11}$$

Define the Hamiltonian as

$$\begin{aligned} H(x_k, u_k, v_k) &= x_k^T Q x_k + x_k^T F x_k + \beta^2 x_k^T x_k \\ +u_k^T R_1 u_k &+ v_k^T R_2 v_k + V(x_{k+1}) - V(x_k). \end{aligned} \tag{12}$$

For linear systems, the value function in (8) can be denoted as

$$V(x_k) = x_k^T P x_k. \tag{13}$$

Based on [27], the necessary conditions for optimal control u_k^* and v_k^* is given by

$$\frac{\partial H(x_k, u_k, v_k)}{\partial u_k} = 0, \quad \frac{\partial H(x_k, u_k, v_k)}{\partial v_k} = 0. \tag{14}$$

Considering (12) and (13), (14) is equivalent to:

$$\begin{bmatrix} R_1 + B^T P B & B^T P D \\ D^T P B & R_2 + D^T P D \end{bmatrix} \begin{bmatrix} u_k^* \\ v_k^* \end{bmatrix} = \begin{bmatrix} -B^T P A \\ -D^T P A \end{bmatrix}$$

Denote $E = B^T P A$, $G = D^T P A$ and the block matrix $\mathscr{M} = \begin{bmatrix} \mathscr{M}_{11} & \mathscr{M}_{12} \\ \mathscr{M}_{21} & \mathscr{M}_{22} \end{bmatrix} = \begin{bmatrix} R_1 + B^T P B & B^T P D \\ D^T P B & R_2 + D^T P D \end{bmatrix}$. Then the optimal control u_k^* and v_k^* can be expressed as $\begin{bmatrix} u_k^* \\ v_k^* \end{bmatrix} = -\mathscr{M}^{-1} \begin{bmatrix} E \\ G \end{bmatrix} x_k$.

Let $\mathscr{M}^{-1} = \mathscr{N}$ be partitioned into the block form as $\mathscr{N} = \begin{bmatrix} \mathscr{N}_{11} & \mathscr{N}_{12} \\ \mathscr{N}_{21} & \mathscr{N}_{22} \end{bmatrix}$. Based on the matrix inversion lemma in [28], $\mathscr{N}$ can be expressed as:

$$\begin{aligned} \mathscr{N}_{11} &= \left(\mathscr{M}_{11} - \mathscr{M}_{12}\mathscr{M}_{22}^{-1}\mathscr{M}_{21}\right)^{-1}, \\ \mathscr{N}_{12} &= \left(\mathscr{M}_{11} - \mathscr{M}_{12}\mathscr{M}_{22}^{-1}\mathscr{M}_{21}\right)^{-1}\mathscr{M}_{12}\mathscr{M}_{22}^{-1}, \\ \mathscr{N}_{21} &= \left(\mathscr{M}_{22} - \mathscr{M}_{21}\mathscr{M}_{11}^{-1}\mathscr{M}_{12}\right)^{-1}\mathscr{M}_{21}\mathscr{M}_{11}^{-1}, \\ \mathscr{N}_{22} &= \left(\mathscr{M}_{22} - \mathscr{M}_{21}\mathscr{M}_{11}^{-1}\mathscr{M}_{12}\right)^{-1}. \end{aligned} \tag{15}$$

Finally, the optimal control can be expressed $u_k^* = K^* x_k$ and $v_k^* = L^* x_k$ with

$$K^* = -\left(\mathscr{N}_{11}E + \mathscr{N}_{12}G\right), \tag{16}$$

$$L^* = -\left(\mathscr{N}_{21}E + \mathscr{N}_{22}G\right). \tag{17}$$

Taking E, G, $\mathscr{M}$ and $\mathscr{N}$ into (16) and (17), then the following can be obtained.

$$\begin{aligned} K^* = &-\Big[B^T P B - B^T P D\big(R_2 + D^T P D\big)^{-1} D^T P B \\ &+ R_1\Big]^{-1}\Big[B^T P A - B^T P D\big(R_2 + D^T P D\big)^{-1} D^T P A\Big], \\ L^* = &-\Big[D^T P D - D^T P B\big(R_1 + B^T P B\big)^{-1} B^T P D \\ &+ R_2\Big]^{-1}\Big[D^T P A - D^T P B\big(R_1 + B^T P B\big)^{-1} B^T P A\Big], \end{aligned}$$

where P is the solution of the algebraic Riccati euqation (ARE)

$$\begin{bmatrix} E \\ G \end{bmatrix}^T \mathscr{M}^{-1} \begin{bmatrix} E \\ G \end{bmatrix} + A^T P A - P + \bar{Q} = 0 \tag{18}$$

As mentioned in Remark 1, the optimal solution to the optimal control problem in (7) and (8) is able to solve the robust stabilization problem only under some specific conditions. The conditions that guarantee the feedback gain K^* in (16) asymptotically stabilizes system (1) is provided as the following theorem.

Theorem 1. *Suppose that there exist a scalar $\varepsilon \succ 0$ satisfies (5) and such that*

$$\varepsilon^{-1} I - P \succ 0. \tag{19}$$

Then the state feedback control $u_k = K^* x_k$ with K^* satisfying (16) can asymptotically stabilize system (1) if it satisfies the following inequality

$$\begin{aligned} &A_c^T\big(P^{-1} - \varepsilon I\big)^{-1} A_c \prec M^T P^{-1} M \\ &+ K^T R_1 K + L^T R_2 L + Q + \beta^2 I \end{aligned} \tag{20}$$

where $M = \left(P^{-1} + B^T R_1^{-1} B + D^T R_2^{-1} D\right)^{-1} A$ and L^* satisfies (17).

Proof. When the feedback gain K^* in (16) is applied to system (1), it can be shown that the function $V(x_k)$ is a Lyapunov function of system (1) if (19) and (20) are satisfied. First, $V(x_k) = x_k^T P x_k \succ 0,\ x_k \neq 0$, since P is the positive definite solution of the ARE (18). Now it remains to show that the time difference $\Delta V(x_k) = V(x_{k+1}) - V(x_k) \prec 0, \forall x_k \neq 0$.

Inserting the feedback gain K^* in (16) into the uncertain closed-loop dynamics (6)

$$x_{k+1} = (A_c + \Delta) x_k \tag{21}$$

where $A_c = A + BK^*$. The time difference of $V(x_k)$ along the state trajectory of (21) is

$$\begin{aligned} \Delta V(x_k) = x_k^T &\left(A_c^T P A_c + \Delta^T P A_c\right. \\ &\left. + A_c^T P \Delta + \Delta^T P \Delta\right) x_k - x_k^T P x_k \end{aligned} \tag{22}$$

Based on (19), the following is true

$$\left(\varepsilon^{-1} I - P\right)^{-1} \succ 0. \tag{23}$$

Using the inequality $a^2 + b^2 \geq 2ab$, and the fact that P and $\left(\varepsilon^{-1} I - P\right)^{-1}$ are positive definite, one can obtain:

$$\begin{aligned} & A_c^T P\big(\varepsilon^{-1} I - P\big)^{-1} P A_c + \varepsilon^{-1} \Delta A^T \Delta A - \Delta^T P \Delta \\ = \ & A_c^T P\big(\varepsilon^{-1} I - P\big)^{-1} P A_c + \Delta^T \left(\varepsilon^{-1} I - P\right) \Delta \\ \geq \ & A_c^T P \Delta + \Delta^T P A_c. \end{aligned} \tag{24}$$

By rearranging items in (24), the following is obtained

$$\begin{aligned} & A_c^T P \Delta + \Delta^T P A_c + \Delta^T P \Delta \\ \leq \ & A_c^T P\big(\varepsilon^{-1} I - P\big)^T P A_c + \varepsilon^{-1} \Delta^T \Delta, \end{aligned} \tag{25}$$

Combining (22) with (25), one can obtain

$$\begin{aligned} \Delta V(x_k) \leq x_k^T &\Big[A_c^T P\big(\varepsilon^{-1} I - P\big)^T P A_c \\ &+ A_c^T P A_c + \varepsilon^{-1} \Delta^T \Delta - P\Big] x_k \end{aligned} \tag{26}$$

Based on the matrix inversion lemma in [28], the following is true

$$P\big(\varepsilon^{-1} I - P\big)^T P + P = \big(P^{-1} - \varepsilon I\big)^{-1}. \tag{27}$$

Considering (27), (26) is equivalent to

$$\begin{aligned} \Delta V(x_k) \ \leq \ & x_k^T \Big[A_c^T\big(P^{-1} - \varepsilon I\big)^{-1} A_c \\ & + \ \varepsilon^{-1} \Delta^T \Delta - P\Big] x_k. \end{aligned} \tag{28}$$

Considering P in (18), then (28) is equivalent to

$$\begin{aligned}\Delta V(x_k) \leq & x_k^T \Big[A_c^T \left(P^{-1}-\varepsilon I\right)^{-1} A_c \\ - & A^T \left(P^{-1}+B^T R_1^{-1} B+D^T R_2^{-1} D\right)^{-1} A \\ + & \varepsilon^{-1} \Delta^T \Delta - \bar{Q}\Big] x_k. \end{aligned} \quad (29)$$

Let $N = P^{-1} + B^T R_1^{-1} B + D^T R_2^{-1} D$, $M = N^{-1} A$, then

$$\begin{aligned} & A^T \left(P^{-1}+B_1^T R_1^{-1} B_1+B_2^T R_2^{-1} B_2\right)^{-1} A \\ = & M^T P^{-1} M + K^T R_1 K + L^T R_2 L \end{aligned} \quad (30)$$

Inserting (30) into (29) yields

$$\begin{aligned}\Delta V(x_k) = & x_k^T \Big[A_c^T \left(P^{-1}-\varepsilon I\right)^{-1} A_c + \varepsilon^{-1} \Delta^T \Delta \\ - & F M^T P^{-1} M - \left(K^{*T}\right) R_1 K^* \\ - & \left(L^{*T}\right) R_2 L^* - Q - \beta^2 I\Big] x_k. \end{aligned} \quad (31)$$

Based on (5), $\Delta V(x_k) \prec 0$ if (20) is satisfied. This completes the proof. □

Remark 2. *Conditions (19) and (20) guarantee the asymptotic stability of system (1) when K^* is applied. Note that the optimal feedback gains K^* and L^* depend on P, the solution of the ARE (18). Even if P is known, the knowledge of the modified system dynamics in (7) is still required for the optimal feedback gain computation.*

4 Off-policy Reinforcement Learning

The condition for the robust stabilization of the optimal feedback gain K^* has been derived in (16). In order to solve the ARE (18), the off-policy reinforcement learning method is developed to derive the optimal feedback gain K^*, L^* in this section. The derivation of the off-policy reinforcement learning algorithm for the discrete-time linear dynamic system (7) is also derived. Through the derivation, it can be seen that the off-policy RL algorithm has the merit that the optimal control problem could be solved without the requirement of the system knowledge.

Suppose the admissible policies $u_k = u(x_k)$ and $u_k = u(x_k)$ are applied to the system (7). The modified nominal system (7) can be rewritten as:

$$x_{k+1} = A^i x_k + B\left(u_k - K^i x_k\right) + D\left(v_k - L^i x_k\right), \quad i = 0, 1, 2, \cdots \quad (32)$$

where $A_i = A + BK^i + DL^i$, $u_k^i = K^i x_k$, $v_k^i = L^i x_k$ and $u^0(x_k) = K^0 x_k$, $v^0(x_k) = L^0 x_k$ are admissible policies. Then the Bellman equation (11) can be rewritten as:

$$\begin{aligned} & V^i(x_k) - V^i(x_{k+1}) = r\left(x_k, u_k^i, v_k^i\right) \\ = & x_k^T Q x_k + x_k^T F x_k + \beta^2 x_k^T x_k \\ + & \left(u_k^i\right)^T R_1 u_k^i + \left(v_k^i\right)^T R_1 v_k^i, \end{aligned} \quad (33)$$

where $V^i(x_k) = x_k^T P^i x_k$. The Taylor series expansion of the value function $\Lambda(x)$ at the state a should be:

$$\begin{aligned}\Lambda(x) = & \Lambda(a) + \left\langle \frac{\partial \Lambda(a)}{\partial a}, (x-a) \right\rangle \\ + & \frac{1}{2}(x-a)^T \frac{\partial^2 \Lambda(a)}{\partial a^2}(x-a). \end{aligned} \quad (34)$$

Considering that $V^i(x_k) = x_k^T P^i x_k$, then (34) is equivalent to:

$$\begin{aligned} V^i(x_k) - V^i(x_{k+1}) = & 2 x_{k+1}^T P^i (x_k - x_{k+1}) \\ & + (x_k - x_{k+1})^T P^i (x_k - x_{k+1}). \end{aligned} \quad (35)$$

By taking (32) into (35), one can obtain:

$$\begin{aligned} & V^i(x_k) - V^i(x_{k+1}) \\ = & x_k^T P^i x_k - x_k^T A_i^T P^i A_i x_k - \left(v_k - L^i x_k\right) D^T P^i x_{k+1} \\ - & \left(v_k - L^i x_k\right) D^T P^i A_i x_k - \left(u_k - K^i x_k\right) B^T P^i x_{k+1} \\ - & \left(u_k - K^i x_k\right) B^T P^i A_i x_k. \end{aligned} \quad (36)$$

By using (11), the following discrete time Lyapunov equation holds:

$$P^i = Q + F + \beta^2 I + K_i^T R_1 K_i + L_i^T R_2 L_i + A_i^T P^i A_i.$$

Therefore, the following holds:

$$\begin{aligned} & x_k^T P^i x_k - x_k^T A_i^T P^i A_i x_k = x_k^T Q x_k + x_k^T F x_k \\ & + \beta^2 x_k^T x_k + x_k^T K_i^T R_1 K_i x_k + x_k^T L_i^T R_2 L_i x_k. \end{aligned} \quad (37)$$

Inserting (37) into (36) gives the off-policy Bellman equation:

$$\begin{aligned} & V^i(x_k) - V^i(x_{k+1}) = x_k^T Q x_k + x_k^T F x_k \\ - & \left(v_k - L^i x_k\right)^T D^T P^i x_{k+1} + x_k^T K_i^T R_1 K_i x_k \\ - & \left(v_k - L^i x_k\right)^T D^T P^i A_i x_k + x_k^T L_i^T R_2 L_i x_k \\ - & \left(u_k - K^i x_k\right)^T B^T P^i x_{k+1} + \beta^2 x_k^T x_k \\ - & \left(u_k - K^i x_k\right)^T B^T P^i A_i x_k. \end{aligned} \quad (38)$$

By using the Kronecker product, the off-policy Bellman equation (38) can be rewritten as:

$$\begin{aligned} & \left(x_k^T \otimes x_k^T\right) vec\left(P^i\right) - \left(x_{k+1}^T \otimes x_{k+1}^T\right) vec\left(P^i\right) \\ + & 2\left[\left(v_k - L^i x_k\right)^T \otimes x_k^T\right] vec\left(D^T P^i A\right) \\ + & \left[\left(v_k - L^i x_k\right)^T \otimes \left(u_k + K^i x_k\right)^T\right] vec\left(D^T P^i B\right) \\ + & \left[\left(v_k - L^i x_k\right)^T \otimes \left(v_k + L^i x_k\right)^T\right] vec\left(D^T P^i D\right) \\ + & 2\left[\left(u_k - K^i x_k\right)^T \otimes x_k^T\right] vec\left(B^T P^i A\right) \\ + & \left[\left(u_k - K^i x_k\right)^T \otimes \left(u_k + K^i x_k\right)^T\right] vec\left(B^T P^i B\right) \\ + & \left[\left(u_k - K^i x_k\right)^T \otimes \left(v_k + L^i x_k\right)^T\right] vec\left(B^T P^i D\right) \\ = & x_k^T Q x_k + x_k^T F x_k + \beta^2 x_k^T x_k + x_k^T K_i^T R_1 K_i x_k \\ + & x_k^T L_i^T R_2 L_i x_k \end{aligned} \quad (39)$$

where the unknown variables collected as

$$\begin{aligned} X^i = \Big[& \left(X_1^i\right)^T \quad \left(X_2^i\right)^T \quad \left(X_3^i\right)^T \\ & \left(X_4^i\right)^T \quad \left(X_5^i\right)^T \quad \left(X_6^i\right)^T \quad \left(X_7^i\right)^T \Big]^T, \end{aligned} \quad (40)$$

with

$$\begin{aligned} X_1^i = & vec\left(P^i\right), X_2^i = vec\left(D^T P^i A\right), \\ X_3^i = & vec\left(D^T P^i B\right) X_4^i = vec\left(D^T P^i D\right), \\ X_5^i = & vec\left(B^T P^i A\right), X_6^i = vec\left(B^T P^i B\right), \\ X_7^i = & vec\left(B^T P^i D\right). \end{aligned}$$

The data collected online in compact form is denoted as:

$$H_k^i = \begin{bmatrix} H_{xx}^{ik} & H_{vx}^{ik} & H_{vu}^{ik} & H_{vv}^{ik} & H_{ux}^{ik} & H_{uu}^{ik} & H_{uv}^{ik} \end{bmatrix},$$

with

$$\begin{aligned}
H_{xx}^{ik} &= \left(x_k^T \otimes x_k^T\right) - \left(x_{k+1}^T \otimes x_{k+1}^T\right), \\
H_{vx}^{ik} &= 2\left[\left(v_k - L^i x_k\right)^T \otimes x_k^T\right], \\
H_{vu}^{ik} &= \left(v_k - L^i x_k\right)^T \otimes \left(u_k + K^i x_k\right)^T, \\
H_{vv}^{ik} &= \left(v_k - L^i x_k\right)^T \otimes \left(v_k + L^i x_k\right)^T, \\
H_{ux}^{ik} &= 2\left[\left(u_k - K^i x_k\right)^T \otimes x_k^T\right], \\
H_{uu}^{ik} &= \left(u_k - K^i x_k\right)^T \otimes \left(u_k + K^i x_k\right)^T, \\
H_{uv}^{ik} &= \left(u_k - K^i x_k\right)^T \otimes \left(v_k + L^i x_k\right)^T.
\end{aligned}$$

Furthermore, denote the online measured utility function

$$\begin{aligned}
r_k^i &= x_k^T Q x_k + x_k^T F x_k + \beta^2 x_k^T x_k + x_k^T K_i^T R_1 K_i x_k \\
&+ x_k^T L_i^T R_2 L_i x_k. \qquad (41)
\end{aligned}$$

The Kronecker product based off-policy Bellman equation (39) can be rewritten in compact form as:

$$H_k^i X^i = r_k. \qquad (42)$$

Note that in (39), there are $N = 3n^2 + m^2 + 3nm$ unknown components. Therefore, at least N data are required to collected in order to solve (39) or (42) by least squares methods. Assumed that $N_1 \geq N$ data are collected as

$$H_{1:N_1} X^i = \begin{bmatrix} H_1^i \\ H_2^i \\ \vdots \\ H_{N_1}^i \end{bmatrix} X^i = \begin{bmatrix} r_1 \\ r_1 \\ \vdots \\ r_{N_1} \end{bmatrix} = r_{1:N_1}. \qquad (43)$$

Therefore, the least squares solution of (43)

$$\hat{X}^i = \left(H_{1:N_1}^T H_{1:N_1}\right)^{-1} H_{1:N_1}^T r_{1:N_1}. \qquad (44)$$

Based on the least squares solution $\hat{X}$ in (44), the feedback gain K^i and L^i are updated as

$$\begin{aligned}
K^{i+1} &= \left[R_1 + X_3^i + X_6^i\left(X_7^i + R_2\right)^{-1} X_5^i\right]^{-1} \\
&\times \left[X_2^i - X_6^i\left(X_7^i + R_2\right)^{-1} X_4^i\right] \\
L^{i+1} &= \left[R_2 + X_7^i - X_5^i\left(R_1 + X_3^i\right)^{-1} X_6^i\right]^{-1} \\
&\times \left[X_4^i + X_5^i\left(R_1 + X_3^i\right)^{-1} X_2^i\right]
\end{aligned}$$

5 Simulation Study

In this section, in order to demonstrate the effectiveness of the presented algorithm in the previous section, a discrete-time rotating inverted pendulum in [29] is considered. The sampling time of the linear discrete-time rotating pendulum model is T = 0:005s. The system dynamics is given as

$$x_{k+1} = (A + \Delta)\, x_k + B u_k, \qquad (45)$$

where

$$\begin{aligned}
A &= \begin{bmatrix} 1.0008 & 0.005 & 0 & 0 \\ 0.3164 & 1.008 & 0 & 0 \\ -0.0004 & 0 & 1 & 0.005 \\ -0.1666 & -0.0004 & 0 & 1 \end{bmatrix}, \\
B &= \begin{bmatrix} -0.0005 & -2.6043 & 0.0101 & 4.0210 \end{bmatrix}^T, \\
\Delta &= \sin(k) * \begin{bmatrix} 0.0007 & 0.0004 & 0.0001 & 0.0001 \\ 0.2764 & 0.1382 & 0.0503 & 0.0377 \\ 0.0020 & 0.0010 & 0.0004 & 0.0003 \\ 0.4291 & 0.2145 & 0.0780 & 0.0585 \end{bmatrix}.
\end{aligned}$$

The uncertainty bound satisfying (5) is given as

$$F = \begin{bmatrix} 4.4061 & 2.2031 & 0.8812 & 0.6609 \\ 2.2031 & 1.1015 & 0.4406 & 0.3305 \\ 0.8812 & 0.4406 & 0.1757 & 0.1320 \\ 0.6609 & 0.3305 & 0.1320 & 0.0983 \end{bmatrix}.$$

and $\varepsilon = 0.001$. For the corresponding optimal control problem, the weight matrix is selected as $Q = I_{4\times 4}$ and $R_1 = R_2 = 1$. The initial state is $x_0 = \begin{bmatrix} 3 & 4 & 5 & 6 \end{bmatrix}^T$. The design parameters α and β that satisfy (20) is selected as $\alpha = 0.01$ and $\beta = 2$. To begin the off-policy reinforcement learning, the initial admissible feedback gains are chosen as

$$\begin{aligned}
K &= \begin{bmatrix} -3.9773 & -0.7095 & 0.1495 & 0.2016 \end{bmatrix}, \\
L &= \begin{bmatrix} -22.7241 & -3.4461 & 3.0583 & 2.1872 \end{bmatrix}.
\end{aligned}$$

Finally, the solution of the ARE in (18) is

$$P = 10^4 \times \begin{bmatrix} 7.5885 & 1.0632 & -0.6042 & -0.6707 \\ 1.0632 & 0.1587 & -0.0889 & -0.0991 \\ -0.6042 & -0.0889 & 0.1519 & 0.0570 \\ -0.6707 & -0.0991 & 0.0570 & 0.0636 \end{bmatrix},$$

and the optimal feedback gain is

$$K^* = \begin{bmatrix} -3.9136 & -0.6983 & 0.1491 & 0.1953 \end{bmatrix} \qquad (46)$$

$$L^* = \begin{bmatrix} -1.8979 & -0.2816 & 0.2426 & 0.1786 \end{bmatrix} \qquad (47)$$

When taking the optimal feedback gain in (46) back to the original system with uncertainty in (45), the system state trajectories is shown in Figure 1. It can be seen from Figure 1 that with the presented optimal control design based method, the robust control problem of the linear dynamic system with bounded uncertainty is solved.

6 Conclusion

This article presents a model-free solution to the robust control problem of the discrete-time linear systems with bounded uncertainty. Inspired by the idea of [12] for continuous-time systems robust control problems, the idea that translating robust control problem to the optimal control problem is adopted in this paper for discrete-time systems. The equivalence that the optimal control law is able to stabilize the original uncertain system is provided. Off-policy RL method is then applied to the transformed optimal control problem, which has two merits. First, off-policy RL method can be used to find the optimal control feedback gain without requiring any knowledge of the system dynamics. Second, the data collected from on-line can be utilized efficiently. A simulation is conducted to validate the robust stability of the proposed algorithm.

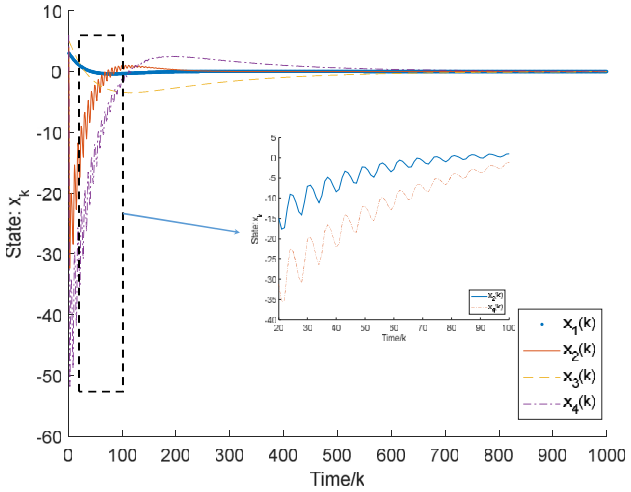

Fig. 1: The systems trajectories x_k.

References

[1] J. Cruz, J. Freudenberg, and D. Looze, "A relationship between sensitivity and stability of multivariable feedback systems," *IEEE Transactions on Automatic Control*, vol. 26, no. 1, pp. 66–74, 1981.

[2] F. Lin, R. D. Brandt, and J. Sun, "Robust control of nonlinear systems: compensating for uncertainty," *International Journal of Control*, vol. 56, no. 6, pp. 1453–1459, 1992.

[3] B. Chen and C. Wong, "Robust linear controller design: time domain approach," *IEEE transactions on automatic control*, vol. 32, no. 2, pp. 161–164, 1987.

[4] H. Ma, Z. Wang, D. Wang, D. Liu, P. Yan, and Q. Wei, "Neural-network-based distributed adaptive robust control for a class of nonlinear multiagent systems with time delays and external noises," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 46, no. 6, pp. 750–758, 2016.

[5] D. Tong, Q. Zhu, W. Zhou, Y. Xu, and J. Fang, "Adaptive synchronization for stochastic t–s fuzzy neural networks with time-delay and markovian jumping parameters," *Neurocomputing*, vol. 117, pp. 91–97, 2013.

[6] D. Tong, W. Zhou, X. Zhou, J. Yang, L. Zhang, and Y. Xu, "Exponential synchronization for stochastic neural networks with multi-delayed and markovian switching via adaptive feedback control," *Communications in Nonlinear Science and Numerical Simulation*, vol. 29, no. 1, pp. 359–371, 2015.

[7] D. Tong, L. Zhang, W. Zhou, J. Zhou, and Y. Xu, "Asymptotical synchronization for delayed stochastic neural networks with uncertainty via adaptive control," *International Journal of Control, Automation and Systems*, vol. 14, no. 3, pp. 706–712, 2016.

[8] F. Lin, *Robust control design: an optimal control approach*. John Wiley & Sons, 2007, vol. 18.

[9] D. Liu, D. Wang, F.-Y. Wang, H. Li, and X. Yang, "Neural-network-based online HJB solution for optimal robust guaranteed cost control of continuous-time uncertain nonlinear systems," *IEEE transactions on cybernetics*, vol. 44, no. 12, pp. 2834–2847, 2014.

[10] X. Yang, D. Liu, Q. Wei, and D. Wang, "Guaranteed cost neural tracking control for a class of uncertain nonlinear systems using adaptive dynamic programming," *Neurocomputing*, vol. 198, pp. 80–90, 2016.

[11] D. Liu, X. Yang, D. Wang, and Q. Wei, "Reinforcement-learning-based robust controller design for continuous-time uncertain nonlinear systems subject to input constraints," *IEEE transactions on cybernetics*, vol. 45, no. 7, pp. 1372–1385, 2015.

[12] D. Wang, D. Liu, H. Li, and H. Ma, "Neural-network-based robust optimal control design for a class of uncertain nonlinear systems via adaptive dynamic programming," *Information Sciences*, vol. 282, pp. 167–179, 2014.

[13] D. Wang, C. Li, D. Liu, and C. Mu, "Data-based robust optimal control of continuous-time affine nonlinear systems with matched uncertainties," *Information Sciences*, vol. 366, pp. 121–133, 2016.

[14] P. J. Werbos, "Approximate dynamic programming for real-time control and neural modeling," *Handbook of intelligent control: Neural, fuzzy, and adaptive approaches*, vol. 15, pp. 493–525, 1992.

[15] D. Kleinman, "On an iterative technique for riccati equation computations," *IEEE Transactions on Automatic Control*, vol. 13, no. 1, pp. 114–115, 1968.

[16] G. N. Saridis and C.-S. G. Lee, "An approximation theory of optimal control for trainable manipulators," *IEEE Transactions on systems, Man, and Cybernetics*, vol. 9, no. 3, pp. 152–159, 1979.

[17] M. Abu-Khalaf and F. L. Lewis, "Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network hjb approach," *Automatica*, vol. 41, no. 5, pp. 779–791, 2005.

[18] D. Vrabie and F. Lewis, "Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems," *Neural Networks*, vol. 22, no. 3, pp. 237–246, 2009.

[19] D. Vrabie, O. Pastravanu, M. Abu-Khalaf, and F. Lewis, "Adaptive optimal control for continuous-time linear systems based on policy iteration," *Automatica*, vol. 45, no. 2, pp. 477–484, 2009.

[20] F. Silvia and R. F. Stengel, "Model-based adaptive critic designs," *Handbook of learning and approximate dynamic programming*, vol. 2, p. 65, 2004.

[21] D. Wang, D. Liu, Q. Wei, D. Zhao, and N. Jin, "Optimal control of unknown nonaffine nonlinear discrete-time systems based on adaptive dynamic programming," *Automatica*, vol. 48, no. 8, pp. 1825–1832, 2012.

[22] Y. Jiang and Z. Jiang, "Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics," *Automatica*, vol. 48, no. 10, pp. 2699–2704, 2012.

[23] B. Luo, H. Wu, and T. Huang, "Off-policy reinforcement learning for control design," *IEEE transactions on cybernetics*, vol. 45, no. 1, pp. 65–76, 2015.

[24] B. Kiumarsi, F. L. Lewis, and Z.-P. Jiang, "H control of linear discrete-time systems: Off-policy reinforcement learning," *Automatica*, vol. 78, pp. 144–152, 2017.

[25] H. Modares, F. L. Lewis, and Z.-P. Jiang, "Tracking control of completely unknown continuous-time systems via off-policy reinforcement learning," *IEEE transactions on neural networks and learning systems*, vol. 26, no. 10, pp. 2550–2562, 2015.

[26] H. Modares, S. P. Nageshrao, G. A. D. Lopes, R. Babuška, and F. L. Lewis, "Optimal model-free output synchronization of heterogeneous systems using off-policy reinforcement learning," *Automatica*, vol. 71, pp. 334–341, 2016.

[27] F. L. Lewis and V. L. Syrmos, *Optimal control*. John Wiley & Sons, 1995.

[28] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.

[29] N. S. Tripathy, I. N. Kar, and K. Paul, "Stabilization of uncertain discrete-time linear system with limited communication," *IEEE Transactions on Automatic Control*, 2017, in Press.