Robust Control for Uncertain Discrete-Time Linear Systems Using Reinforcement Learning With Discount Factor

Yuntian Ding, Yuxiao Yang, Zhilian Yan, Weipeng Tai

***Abstract*—This paper focuses on robust control for uncertain discrete-time linear systems. We transform the robust control issure into an optimal control issure for an auxiliary system. The desired control gains are obtained by solving coupled algebraic Riccati equations. We propose a model-free off-policy reinforcement learning algorithm that incorporates a discount factor to get an approximate solution to the equation, using only measured data and eliminating the need for information about system dynamics. This algorithm requires adding probing noise to control inputs to maintain persistent excitation condition. We show that the probing noise does not introduce bias in the solution of the Bellman equation. The effectiveness of the proposed algorithm is verified through a simulation example, which analyzes the impact of the discount factor and the probing noise.**

***Index Terms*—Reinforcement learning, discount factor, model free control, off-policy algorithm, parameter uncertainty.**

I. Introduction

DISCRETE-TIME linear systems (DLSs) are typical control systems that process signals sampled at discrete intervals, as opposed to continuously evolving signals. These systems often use difference equations and state space equations to describe their mathematical models. They are widely used in digital signal processing, control theory, communication systems, and computer science [1]–[4].

Uncertainty exists widely in actual control systems, which can result in system performance degradation and, in extreme cases, system instability [5]–[10]. Over the past few decades, many robust control methods have been proposed to guarantee the performance of uncertain DLSs. For example, De Souza and Coutinho [11] examined periodic DLSs subject to delay state and polyhedral parameter uncertainty and proposed a criterion for robust stability. Morais *et al.* [12] focuses on Markov jump DLSs with uncertain transition probabilities and develops a reduced-order dynamic compensation control scheme. Marcela *et al.* [13] dealt with guaranteed cost control of DLSs with uncorrelated block diagonal structural parameter uncertainty and developed conditions for state feedback and output feedback, respectively. Recently, Jiang *et al.* [14] studied the consensus of uncertain DLS modelling multi-agent systems and proposed two distributed adaptive control algorithms.

The methods presented in [11]–[14] rely on accurate system models. However, accurate mathematical models are difficult to obtain in practical situations, limiting the use of these methods. Thanks to the advantages of machine learning, model-free reinforcement learning (RL) methods for solving robust control problems of uncertain DLSs have become a widely discussed topic in the control community. Typical model-free RL methods are divided into on-policy RL and off-policy RL methods. Unlike on-policy RL methods, off-policy RL methods can learn from a wider data distribution and generally exhibit higher sample efficiency [15]. Taking this into consideration, Yang *et al.* [16] explored the robust stabilization of DLSs with bounded and mismatched uncertainty, proposing a model-free off-policy RL method that does not require knowledge of system dynamics.

In this paper, we investigate the issue of RL-based robust control for DLSs with mismatching uncertainty. Unlike the work of [16], we introduce a discount factor into the cost function to speed up learning and reduce control costs. We transform the robust control problem into an optimal control problem by constructing an auxiliary system. Then, we propose a model-free off-policy RL algorithm to solve the optimal control problem, incorporating probing noise and the least-squares method. It is theoretically proven that the probing noise does not bias the solution after introducing the discount factor. Finally, we validate the proposed theoretical results through several simulation examples.

The rest of this paper is organized as follows: In Section II, the relevant description of the uncertain DLSs is given, and the robust control problem is transformed into an optimal control problem by constructing auxiliary systems. In Section III, the model-free off-policy RL method for solving the optimal control problem of auxiliary system is introduced, and the effect of probing noise on this method is investigated. In Section IV, simulation results under different discount factors and different probing noise are given to verify the impact of the discount factor and the probing noise.

II. Preliminaries

A. Uncertain System

In this paper, a class of uncertain DLSs is expressed by the following equation as [17]:

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

Manuscript received June 16, 2024; revised October 31, 2024.

Yuntian Ding is a postgraduate student at the Research Institute of Information Technology, Anhui University of Technology, Ma'anshan 243032, China (e-mail: ytding@ahut.edu.cn).

Yuxiao Yang is a college student at the School of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, Suzhou 215000, China (e-mail: yuxiao.yang21@student.xjtlu.edu.cn)

Zhilian Yan is a lecturer of the School of Electrical & Information Engineering, Anhui University of Technology, Ma'anshan 243032, China (e-mail: zlyan@ahut.edu.cn).

Weipeng Tai is a full professor at the School of Computer Science and Technology, Anhui University of Technology, Ma'anshan 243032, China (e-mail: taiweipeng@ahut.edu.cn).

with the nominal system

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

where $x_k \in R^n$ is the state, $u_k \in R^m$ is the control input. $A \in R^{n\times n}$ and $B \in R^{n\times m}$ are system parameters and are constant matrices. $\Delta A(p) \in R^{n\times n}$ is an unknown matrix, which represents the uncertainty of the system through bounded variation of p. The parameter p belongs to a preset bounded set Ω.

System uncertainty can be classified into two types, matching uncertainty and mismatching uncertainty [17]. The uncertainty in the system is matching uncertainty when the uncertainty matrix $\Delta A(p)$ can be expressed as

$$\Delta A(p) = B\phi_A(p), \quad \forall p \in \Omega \tag{3}$$

where $\phi_A(p)$ is an uncertain perturbation related to parameter p. For mismatching uncertainty, $\Delta A(p)$ cannot be represented in the form of (3). Moreover, mismatching uncertainty can be decomposed into two parts, such as

$$\begin{aligned}\Delta A(p) &= S\phi_A(p)\\ &= BB^+S\phi_A(p) + (I - BB^+)S\phi_A(p), \quad \forall p \in \Omega\end{aligned} \tag{4}$$

where $BB^+S\phi_A(p)$ is matched and $(I - BB^+)S\phi_A(p)$ is mismatched. In equation (4), $B^+ = (B^\top B)^{-1}B^\top$ is the pseudo-inverse of the matrix B [18], S $\neq$ B, S is the known matrix, $\Delta A(p)$ is the uncertain perturbation. The perturbation $\phi_A(p)$ is upper bounded by positive semide-finite matrix F and defined as

$$\varepsilon^{-1}\phi_A^T(p)\phi_A(p) \leq F, \quad \forall p \in \Omega \tag{5}$$

where ε is a positive constant. To effectively control system (1), it is necessary to pose the following robust control issure.

B. Robust Control Issure

The robust control issure can be formulated as finding a suitable control law $u_k = Kx_k$ in terms of state feedback [19]–[24] so that the uncertain system

$$x_{k+1} = (A + BK)x_k + \Delta A(p)x_k \tag{6}$$

can be asymptotically stable for $\forall p \in \Omega$.

To design a robust control law $u_k = Kx_k$ that can make system (1) asymptotically stable, we can construct an auxiliary system by introducing additional term Dv_k, thereby converting the robust control issure into an optimal control issure. We can then use optimal control methods to solve for the feedback gain K. The auxiliary system can be obtained [16] as

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

where $D = \alpha(I - BB^+)S \in R^{n\times r}$ and α is a positive constant.

Remark 1. Based on the discussion above, robust control issure for uncertain system (1) can be transformed as an optimal control issure for an auxiliary system (7). The control input u_k in auxiliary system (7) directly affects uncertain system (1). It is important to note that the control input v_k only appears in auxiliary system (7) and does not directly influence uncertain system (1). In fact, v_k can be regarded as a virtual input.

C. Optimal Control Approach

In this subsection, we apply the optimal control method to address the robust control issure for uncertain system (1). To aid in the discussion the following definition is provided.

Definition 1. (Admissible Control) [25]: For auxiliary system (7), the feedback control laws $u(x_k)$ and $v(x_k)$ are admissible if the following conditions are satisfied:

1) $u(x_k)$ and $v(x_k)$ are continuous;
2) $u(0) = v(0) = 0$;
3) $u(x_k)$ and $v(x_k)$ can stabilize system (7).

The discount factor is crucial for balancing the weight of current and future rewards. Existing research shows that The more minor the discount factor, the faster the algorithm converges and the lower the control cost [26], [27]. Motivated by this conclusion, we introduce the discount factor into the robust optimal control of uncertain DLSs. For system (7), the value function can be defined as

$$\begin{aligned}&V(x_k)\\ &= \sum_{j=k}^{\infty}\{x_j^\top(\beta^2 I + Q + F)x_j + u_j^\top R_1 u_j + v_j^\top R_2 v_j\}\end{aligned} \tag{8}$$

where Q ≥ 0, $R_1 > 0$, $R_2 > 0$ and β is a positive constant. After introducing the discount factor, the discounted value function is defined as

$$\begin{aligned}&V_\gamma(x_k)\\ &= \sum_{j=k}^{\infty}\gamma^{j-k}\{x_j^\top(\beta^2 I + Q + F)x_j + u_j^\top R_1 u_j + v_j^\top R_2 v_j\}\end{aligned} \tag{9}$$

where γ is the discount factor $\in (0, 1]$. When $\gamma = 1$, value function (8) is equivalent to discount value function (9).

To simplify the equation, let $\bar{Q} = \beta^2 I + F + Q$. By (9), under the admissible control laws $u_k = Kx_k$ and $v_k = Lx_k$, the discounted value function is defined as

$$V_\gamma(x_k) = \sum_{j=k}^{\infty}\gamma^{j-k}\left\{x_j^\top\bar{Q}x_j + u_j^\top R_1 u_j + v_j^\top R_2 v_j\right\}. \tag{10}$$

The goal of optimal control is to find optimal feedback control laws

$$u_k^* = K^* x_k, \quad v_k^* = L^* x_k \tag{11}$$

which minimize discounted value function (10), such as

$$\begin{aligned}&V_\gamma^*(x_k)\\ &= \min_{u_k, v_k}\sum_{j=k}^{\infty}\gamma^{j-k}\left\{x_j^\top\bar{Q}x_j + u_j^\top R_1 u_j + v_j^\top R_2 v_j\right\}.\end{aligned} \tag{12}$$

Transform the expression of discounted value function (10) into

$$\begin{aligned}V_\gamma(x_k) = &\sum_{j=k+1}^{\infty}\gamma^{j-k}\{x_j^\top\bar{Q}x_j + u_j^\top R_1 u_j + v_j^\top R_2 v_j\}\\ &+ x_k^\top\bar{Q}x_k + u_k^\top R_1 u_k + v_k^\top R_2 v_k.\end{aligned} \tag{13}$$

Using discounted value function (13), the following Bellman equation (BE) can be obtained

$$V_\gamma(x_k) = x_k^\top\bar{Q}x_k + u_k^\top R_1 u_k + v_k^\top R_2 v_k + \gamma V_\gamma(x_{k+1})$$

$$= x_k^\top P x_k. \tag{14}$$

Then, BE (14) is equivalent to

$$\begin{aligned} V_\gamma(x_k) = {} & x_k^\top \bar{Q} x_k + u_k^\top R_1 u_k + v_k^\top R_2 v_k \\ & + \gamma x_{k+1}^\top P x_{k+1}. \end{aligned} \tag{15}$$

Therefore, the Hamiltonian function of system (7) is defined as

$$\begin{aligned} H(x_k, u_k, v_k) = {} & x_k^\top \bar{Q} x_k + u_k^\top R_1 u_k + v_k^\top R_2 v_k \\ & + \gamma x_{k+1}^\top P x_{k+1} - x_k^\top P x_k \end{aligned} \tag{16}$$

Based on [25], optimal control laws u_k^* and v_k^* satisfies the following conditions

$$\frac{H(x_k, u_k, v_k)}{\partial u_k} = 0, \quad \frac{H(x_k, u_k, v_k)}{\partial v_k} = 0 \tag{17}$$

which are equivalent to

$$\begin{bmatrix} R_1 + \gamma B^\top P D & \gamma B^\top P D \\ \gamma D^\top P B & R_2 + \gamma D^\top P D \end{bmatrix} \begin{bmatrix} u_k^* \\ v_k^* \end{bmatrix} = -\gamma \begin{bmatrix} B^\top P A \\ D^\top P A \end{bmatrix} x_k. \tag{18}$$

Define the following variables:

$$\begin{aligned} \xi &= B^\top P A \\ \varrho &= D^\top P A \\ S &= \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \\ &= \begin{bmatrix} R_1 + B^\top P B & \gamma B^\top P D \\ \gamma D^\top P B & R_2 + \gamma D^\top P D \end{bmatrix}. \end{aligned}$$

Then, u_k^* and v_k^* can be expressed as

$$\begin{bmatrix} u_k^* \\ v_k^* \end{bmatrix} = -S^{-1} \begin{bmatrix} \xi \\ \varrho \end{bmatrix} x_k. \tag{19}$$

Let

$$Z = S^{-1} = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix}. \tag{20}$$

Then, according to the matrix inversion lemma [18], each component of the matrix can be expressed as

$$\begin{aligned} Z_{11} &= \left(S_{11} - S_{12} S_{22}^{-1} S_{21}\right)^{-1} \\ Z_{12} &= -\left(S_{11} - S_{12} S_{22}^{-1} S_{21}\right)^{-1} S_{12} S_{22}^{-1} \\ Z_{21} &= -\left(S_{22} - S_{21} S_{11}^{-1} S_{12}\right)^{-1} S_{21} S_{11}^{-1} \\ Z_{22} &= \left(S_{22} - S_{21} S_{11}^{-1} S_{12}\right)^{-1}. \end{aligned}$$

The optimal control laws can be expressed as (11), where K^* and L^* satisfy

$$K^* = -\left(Z_{11}\xi + Z_{12}\varrho\right), \tag{21}$$

$$L^* = -\left(Z_{21}\xi + Z_{22}\varrho\right), \tag{22}$$

which are equivalent to

$$\begin{aligned} K^* = {} & -[R_1 + \gamma B^\top P B - \gamma^2 B^\top P D \theta_2^{-1} D^\top P B]^{-1} \\ & \times [\gamma B^\top P A - \gamma^2 B^\top P D \theta_2^{-1} D^\top P A] \end{aligned} \tag{23}$$

$$\begin{aligned} L^* = {} & -[R_2 + \gamma D^\top P D - \gamma^2 D^\top P B \theta_1^{-1} B^\top P D]^{-1} \\ & \times [\gamma D^\top P A - \gamma^2 D^\top P B \theta_1^{-1} B^\top P A] \end{aligned} \tag{24}$$

respectively, and $\theta_1 = R_1 + \gamma B^\top P B, \ \theta_2 = R_2 + \gamma D^\top P D$. According to equation (17), control laws u_k^* and v_k^* satisfy

$$0 = \min_{u_k, v_k} H\left(x_k, u_k, v_k\right) = H(x_k, u_k^*, v_k^*)$$

which can be written as

$$\begin{aligned} 0 = {} & \begin{bmatrix} u_k^* \\ v_k^* \end{bmatrix}^\top \begin{bmatrix} R_1 + \gamma B^\top P B & \gamma B^\top P D \\ \gamma D^\top P B & R_2 + \gamma D^\top P D \end{bmatrix} \begin{bmatrix} u_k^* \\ v_k^* \end{bmatrix} \\ & + \begin{bmatrix} u_k^* \\ v_k^* \end{bmatrix}^\top \begin{bmatrix} \gamma B^\top P A \\ \gamma D^\top P A \end{bmatrix} x_k + x_k^\top \begin{bmatrix} \gamma B^\top P A \\ \gamma D^\top P A \end{bmatrix}^\top \begin{bmatrix} u_k^* \\ v_k^* \end{bmatrix} \\ & + x_k^\top (\gamma A^\top P A - P) x_k + x_k^\top \bar{Q} x_k. \end{aligned} \tag{25}$$

Substituting (19) into (25) yields

$$\begin{aligned} 0 = {} & \begin{bmatrix} \xi \\ \varrho \end{bmatrix}^\top Z^\top \begin{bmatrix} R_1 + \gamma B^\top P B & \gamma B^\top P D \\ \gamma D^\top P B & R_2 + \gamma D^\top P D \end{bmatrix} Z \begin{bmatrix} \xi \\ \varrho \end{bmatrix} \\ & - \begin{bmatrix} \xi \\ \varrho \end{bmatrix}^\top Z^\top \begin{bmatrix} \gamma B^\top P A \\ \gamma D^\top P A \end{bmatrix} - \begin{bmatrix} \gamma B^\top P A \\ \gamma D^\top P A \end{bmatrix}^\top Z \begin{bmatrix} \xi \\ \varrho \end{bmatrix} \\ & + \gamma A^\top P A - P + \bar{Q}. \end{aligned} \tag{26}$$

Substituting (21) and (22) into (26) yields

$$\begin{aligned} 0 = {} & \begin{bmatrix} K^* \\ L^* \end{bmatrix}^\top \begin{bmatrix} R_1 + \gamma B^\top P B & \gamma B^\top P D \\ \gamma D^\top P B & R_2 + \gamma D^\top P D \end{bmatrix} \begin{bmatrix} K^* \\ L^* \end{bmatrix} \\ & + \begin{bmatrix} K^* \\ L^* \end{bmatrix}^\top \begin{bmatrix} \gamma B^\top P A \\ \gamma D^\top P A \end{bmatrix} + \begin{bmatrix} \gamma B^\top P A \\ \gamma D^\top P A \end{bmatrix}^\top \begin{bmatrix} K^* \\ L^* \end{bmatrix} \\ & + \gamma A^\top P A - P + \bar{Q} \end{aligned} \tag{27}$$

with K^* defined as in (23) and L^* defined as in (24).

Equation (27) represents the coupled algebraic Riccati equations (CARE). In this section, by constructing auxiliary system (7), the robust control issure of finding appropriate feedback gain K is transformed into an optimal control issure of finding optimal feedback gains K^* and L^*, which can be obtained by solving CARE.

III. Off-Policy Reinforcement Learning

Since CARE (27) is a nonlinear equation of P, it isn't easy to solve it directly. Therefore, this section proposes a model-free off-policy RL algorithm to obtain an approximate solution to P in CARE (27). This algorithm does not require system dynamics information. Furthermore, it is demonstrated that when probing noise is added to the control inputs, this algorithm obtains unbiased results.

A. *Model-Free Off-policy RL Algorithm*

By using the off-policy RL method, system (7) is written as

$$x_{k+1} = A_k x_k + B(u_k - K^j x_k) + D(v_k - L^j x_k) \tag{28}$$

where $A_k = A + BK^j + DL^j$. In equation (28), $u_k^j = K^j x_k$ and $v_k^j = L^j x_k$ are iterative control policies, and u_k and v_k are behavior policies.

During the iterative process, BE (14) can be expressed as

$$\begin{aligned} x_{k+1}^\top P^{j+1} x_{k+1} = {} & x_k^\top \bar{Q} x_k + (u_k^{j+1})^\top R_1 u_k^{j+1} \\ & + (v_k^{j+1})^\top R_2 v_k^{j+1} + \gamma x_{k+1}^\top P^{j+1} x_{k+1} \end{aligned} \tag{29}$$

and the discounted value function can be expressed as

$$V_\gamma^{j+1}(x_{k+1}) = x_k^\top P^{j+1} x_k. \tag{30}$$

Using (28), (29) and (30), BE (14) can be expressed as

$$\begin{aligned} & V_\gamma^{j+1}(x_k) - \gamma V_\gamma^{j+1}(x_{k+1}) \\ &= -\gamma x_k^\top A_k^\top P^{j+1} A_k x_k + x_k^\top P^{j+1} x_k \\ &\quad - \gamma(u_k - K^j x_k)^\top B^\top P^{j+1} x_{k+1} \\ &\quad - \gamma(u_k - K^j x_k)^\top B^\top P^{j+1} A_k x_k \\ &\quad - \gamma(v_k - L^j x_k)^\top D^\top P^{j+1} x_{k+1} \\ &\quad - \gamma(v_k - L^j x_k)^\top D^\top P^{j+1} A_k x_k. \end{aligned} \tag{31}$$

Using (30) and (31), one can obtain the off-policy BE

$$\begin{aligned} & x_k^\top P^{j+1} x_k - \gamma x_{k+1}^\top P^{j+1} x_{k+1} \\ &= x_k^\top \bar{Q} x_k + x_k^\top (K^j)^\top P^{j+1} K^j x_k \\ &\quad + x_k^\top (L^j)^\top P^{j+1} L^j x_k \\ &\quad - \gamma(u_k - K^j x_k)^\top B^\top P^{j+1} x_{k+1} \\ &\quad - \gamma(u_k - K^j x_k)^\top B^\top P^{j+1} A_k x_k \\ &\quad - \gamma(v_k - L^j x_k)^\top D^\top P^{j+1} x_{k+1} \\ &\quad - \gamma(v_k - L^j x_k)^\top D^\top P^{j+1} A_k x_k. \end{aligned} \tag{32}$$

The value of P^{j+1} is solved by off-policy BE (32), and then the iterative control policies are updated as

$$\begin{aligned} u_k^{j+1} &= K^{j+1} x_k \\ &= (R_1 + \gamma B^\top P^{j+1} B \\ &\quad - \gamma^2 B^\top P^{j+1} D \theta_2^{-1} D^\top P^{j+1} B)^{-1} \\ &\quad \times (\gamma B^\top P^{j+1} A \\ &\quad - \gamma^2 B^\top P^{j+1} D \theta_2^{-1} D^\top P^{j+1} A) x_k \end{aligned} \tag{33}$$

$$\begin{aligned} v_k^{j+1} &= L^{j+1} x_k \\ &= (R_2 + \gamma D^\top P^{j+1} D \\ &\quad - \gamma^2 D^\top P^{j+1} B \theta_2^{-1} B^\top P^{j+1} D)^{-1} \\ &\quad \times (\gamma D^\top P^{j+1} A \\ &\quad - \gamma^2 D^\top P^{j+1} B \theta_2^{-1} B^\top P^{j+1} A) x_k. \end{aligned} \tag{34}$$

Based on the Kronecker product, off-policy BE (32) can be written as

$$\begin{aligned} & (x_k^\top \otimes x_k^\top) vec(P^{j+1}) - \gamma(x_{k+1}^\top \otimes x_{k+1}^\top) vec(P^{j+1}) \\ & + 2\gamma[(u_k - K^j x_k)^\top \otimes x_k^\top] vec(B^\top P^{j+1} A) \\ & + \gamma[(u_k - K^j x_k)^\top \otimes (u_k + K^j x_k)^\top] vec(B^\top P^{j+1} B) \\ & + \gamma[(u_k - K^j x_k)^\top \otimes (v_k + L^j x_k)^\top] vec(B^\top P^{j+1} D) \\ & + 2\gamma[(v_k - L^j x_k)^\top \otimes x_k^\top] vec(D^\top P^{j+1} A) \\ & + \gamma[(v_k - L^j x_k)^\top \otimes (u_k + K^j x_k)^\top] vec(D^\top P^{j+1} B) \\ & + \gamma[(v_k - L^j x_k)^\top \otimes (v_k + L^j x_k)^\top] vec(D^\top P^{j+1} D) \\ = {} & x_k^\top \bar{Q} x_k + x_k^\top (K^j)^\top R_1 K^j x_k + x_k^\top (L^j)^\top R_2 L^j x_k \end{aligned} \tag{35}$$

which is a scalar equation with $n^2 + m^2 + r^2 + 2mr + n(m+r)$ unknown parameters. This means that at least $n^2 + m^2 + r^2 + 2mr + n(m+r)$ data are required to solve iteratively using least squares (LS). Given s$(s > n^2 + m^2 + r^2 + 2mr + n(m+r))$ independent system states $x_{k,(1)}, x_{k,(2)}, \ldots, x_{k,(s)}$, one defines

$$\phi^j = \begin{bmatrix} r(x_{k,(1)}, K^j, L^j) \\ r(x_{k,(2)}, K^j, L^j) \\ \vdots \\ r(x_{k,(s)}, K^j, L^j) \end{bmatrix} \tag{36}$$

with

$$\begin{aligned} & r(x_{k,(j)}, K^j, L^j) \\ &= x_{k,(j)}^\top \bar{Q} x_{k,(j)} + x_{k,(j)}^\top (K^j)^\top R_1 K^j x_{k,(j)} \\ &\quad + x_{k,(j)}^\top (L^j)^\top R_2 L^j j x_{k,(j)} \end{aligned}$$

and

$$\psi^j = \begin{bmatrix} h_{(xx)1} & h_{(xu)1} & h_{(uu)1} & h_{(xv)1} & h_{(uv)1} & h_{(vv)1} \\ h_{(xx)2} & h_{(xu)2} & h_{(uu)2} & h_{(xv)2} & h_{(uv)2} & h_{(vv)2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ h_{(xx)s} & h_{(xu)s} & h_{(uu)s} & h_{(xv)s} & h_{(uv)s} & h_{(vv)s} \end{bmatrix} \tag{37}$$

with

$$\begin{aligned} h_{(xx)j} &= x_{k,(j)}^\top \otimes x_{k,(j)} - \gamma[x_{k,(j+1)}^\top \otimes x_{k,(j+1)}] \\ h_{(xu)j} &= 2\gamma[(v_k - L^j x_{k,(j)})^\top \otimes x_{k,(j)}^\top] \\ h_{(uu)j} &= \gamma[(u_k - K^j x_{k,(j)})^\top \otimes (u_k + K^j x_{k,(j)})^\top] \\ h_{(xv)j} &= 2\gamma[(v_k - K^j x_{k,(j)})^\top \otimes x_{k,(j)}^\top] \\ h_{(uv)j} &= \gamma[(v_k - L^j x_{k,(j)})^\top \otimes (u_k + K^j x_{k,(j)})^\top \\ &\qquad + (u_k - K^j x_{k,(j)})^\top \otimes (v_k + L^j x_{k,(j})^\top] \\ h_{(vv)j} &= \gamma[(v_k - L^j x_{k,(j)})^\top \otimes (v_k + L^j x_{k,(j)})^\top]. \end{aligned}$$

Then the unknown variables are defined in the form of a vector as

$$X^{j+1} = \left[vec(X_1^{j+1}), vec(X_2^{j+1}), \ldots, vec(X_6^{j+1}) \right] \tag{38}$$

with

$$\begin{aligned} X_1^{j+1} &= P^{j+1}, & X_2^{j+1} &= D^\top P^{j+1} A \\ X_3^{j+1} &= D^\top P^{j+1} B, & X_4^{j+1} &= D^\top P^{j+1} D \\ X_5^{j+1} &= B^\top P^{j+1} A, & X_6^{j+1} &= B^\top P^{j+1} B. \end{aligned}$$

Finally, off-policy BE (32) can be formulated by Kronecker product as

$$\psi^j X^{j+1} = \phi^{j+1}. \tag{39}$$

Thus, the LS solution is given by

$$X^{j+1} = ((\psi^j)^\top \psi)^{-1} (\psi^j)^\top \phi^j. \tag{40}$$

Based on the LS solution X^{j+1} in (40), the iterative control policies K^{j+1} and L^{j+1} can be obtained as

$$\begin{aligned} K^{j+1} &= -[R_1 + X_6^{j+1} + X_3^{j+1}(R_2 + X_4^{j+1})^{-1} X_3^{j+1}]^{-1} \\ &\quad \times [X_5^{j+1} - X_3^{j+1}(R_2 + X_4^{j+1})^{-1} X_2^{j+1}] \end{aligned} \tag{41}$$

$$\begin{aligned} L^{j+1} &= -[R_2 + X_4^{j+1} + X_3^{j+1}(R_1 + X_6^{j+1})^{-1} X_3^{j+1}]^{-1} \\ &\quad \times [X_2^{j+1} - X_3^{j+1}(R_2 + X_6^{j+1})^{-1} X_5^{j+1}]. \end{aligned} \tag{42}$$

From (39), it is evident that to guarantee the uniqueness of the solution X^{j+1}, the matrix ψ^{j+1} must be full rank, which is equivalent to satisfying the persistent excitation condition.

Definition 2. (Persistent Excitation): [28] A vector sequence $\boldsymbol{\eta} = [\eta_1, \eta_2, \ldots, \eta_q]^\top$, where q is termed as persistently exciting if there exists a constant $\beta > 0$ such that

$$\sum_{i=k+1}^{k+l} \eta_i \eta_i^\top \geq \beta I \tag{43}$$

and if $q < l$, condition (43) does not hold.

Algorithm 1: Model-Free Off-Policy RL Algorithm

```
Set admissible behavior policies $\hat{u}_k$ and $\hat{v}_k$.
Initialize the iteration number $j$ to 0.
Initialize the iteration policies $u_k^0 = u_k$ and $v_k^0 = v_k$.
Initialize the predetermined error bounds $\delta_1$ and $\delta_2$.
for $j$ do
    Solve $X^{j+1}$ by LS equation (40).
    Update iteration feedback gain $K^{j+1}$ using (41).
    Update iteration feedback gain $L^{j+1}$ using (42).
    if $\left\|K^{j+1} - K^j\right\| < \delta_1$ and $\left\|L^{j+1} - L^j\right\| < \delta_2$ then
        Stop.
    else
        $j = j + 1$.
    end
end
```

To satisfy the persistent excitation condition, probing noise must be added into the control inputs [29]. In this case, the behavior policies u_k and v_k become

$$\hat{u}_k = u_k + e_{k1} \tag{44}$$

$$\hat{v}_k = v_k + e_{k2} \tag{45}$$

where e_{k1} and e_{k2} are probing noise. Based on the above discussion, at the end of this subsection, the model-free off-policy RL algorithm is shown in Algorithm 1.

B. Effect of Probing Noise

This subsection will show that the addition of probing noise does not bias the solution when solving off-policy BE (32).

Theorem 1. Set P^{j+1} is the solution of off-policy BE (32) without adding probing noise, and $\hat{P}^{j+1}$ is the solution of off-policy BE (32) with probing noise. Then $P^{j+1} = \hat{P}^{j+1}$.

Proof: Off-policy BE (32) for control inputs $\hat{u}_k$ and $\hat{v}_k$ is

$$\begin{aligned}
&x_k^\top \hat{P}^{j+1} x_k - \gamma(A_k x_k + B(\hat{u}_k - K^j x_k) \\
&+ D(\hat{v}_k - L^j x_k))^\top \\
&\times \hat{P}^{j+1}(A_k x_k + B(\hat{u}_k - K^j x_k) + D(\hat{v}_k - L^j x_k)) \\
&= x_k^\top \bar{Q} x_k + x_k^\top (K^j)^\top \hat{P}^{j+1} K^j x_k + x_k^\top (L^j)^\top \hat{P}^{j+1} L^j x_k \\
&- \gamma(\hat{u}_k - K^j x_k)^\top B^\top \hat{P}^{j+1} \\
&\times (A_k x_k + B(\hat{u}_k - K^j x_k) + D(\hat{v}_k - L^j x_k)) \\
&- \gamma(\hat{u}_k - K^j x_k)^\top B^\top \hat{P}^{j+1} A_k x_k \\
&- \gamma(\hat{v}_k - L^j x_k)^\top D^\top \hat{P}^{j+1} \\
&\times (A_k x_k + B(u_k - K^j x_k) + D(\hat{v}_k - L^j x_k)) \\
&- \gamma(\hat{v}_k - L^j x_k)^\top D^\top \hat{P}^{j+1} A_k x_k.
\end{aligned} \tag{46}$$

Substituting (44) and (45) into (46) yields

$$\begin{aligned}
&x_k^\top \hat{P}^{j+1} x_k - \gamma(A_k x_k + B(u_k + e_{k1} - K^j x_k) \\
&+ D(v_k + e_{k2} - L^j x_k))^\top \hat{P}^{j+1} \\
&\times (A_k x_k + B(u_k + e_{k1} - K^j x_k) \\
&+ D(v_k + e_{k2} - L^j x_k)) \\
&= x_k^\top \bar{Q} x_k + x_k^\top (K^j)^\top \hat{P}^{j+1} K^j x_k \\
&+ x_k^\top (L^j)^\top \hat{P}^{j+1} L^j x_k \\
&- \gamma(u_k + e_{k1} - K^j x_k)^\top B^\top \hat{P}^{j+1} \\
&\times (A_k x_k + B(u_k + e_{k1} - K^j x_k) \\
&+ D(v_k + e_{k2} - L^j x_k)) \\
&- \gamma(u_k + e_{k1} - K^j x_k)^\top B^\top \hat{P}^{j+1} A_k x_k \\
&- \gamma(v_k + e_{k2} - L^j x_k)^\top D^\top \hat{P}^{j+1} \\
&\times (A_k x_k + B(u_k + e_{k1} - K^j x_k) \\
&+ D(v_k + e_{k2} - L^j x_k)) \\
&- \gamma(v_k + e_{k2} - L^j x_k)^\top D^\top \hat{P}^{j+1} A_k x_k.
\end{aligned} \tag{47}$$

Substituting (28) into (47) yields

$$\begin{aligned}
&x_k^\top \hat{P}^{j+1} x_k - \gamma(x_{k+1} + Be_{k1} + De_{k2})^\top \\
&\times \hat{P}^{j+1}(x_{k+1} + Be_{k1} + De_{k2}) \\
&= x_k^\top \bar{Q} x_k + x_k^\top (K^j)^\top \hat{P}^{j+1} K^j x_k \\
&+ v_k^\top (L^j)^\top \hat{P}^{j+1} L^j v_k \\
&- \gamma(u_k + e_{k1} - K^j x_k)^\top B^\top \hat{P}^{j+1} \\
&\times (x_{k+1} + Be_{k1} + De_{k2}) \\
&- \gamma(u_k + e_{k1} - K^j x_k)^\top B^\top \hat{P}^{j+1} A_k x_k \\
&- \gamma(v_k + e_{k2} - L^j x_k)^\top D^\top \hat{P}^{j+1} \\
&\times (x_{k+1} + Be_{k1} + De_{k2}) \\
&- \gamma(v_k + e_{k2} - L^j x_k)^\top D^\top \hat{P}^{j+1} A_k x_k.
\end{aligned} \tag{48}$$

Expanding the terms on both sides of (48) yields

$$\begin{aligned}
&x_k^\top \hat{P}^{j+1} x_k - \gamma x_{k+1}^\top \hat{P}^{j+1} x_{k+1} \\
&- 2\gamma x_{k+1} \hat{P}^{j+1}(Be_{k1} + De_{k2}) \\
&- \gamma(Be_{k1} + De_{k2})^\top \hat{P}^{j+1}(Be_{k1} + De_{k2}) \\
&= x_k^\top \bar{Q} x_k + x_k^\top (K^j)^\top \hat{P}^{j+1} K^j x_k \\
&+ x_k^\top (L^j)^\top \hat{P}^{j+1} L^j x_k \\
&- \gamma(u_k - K^j x_k)^\top B^\top \hat{P}^{j+1} x_{k+1} \\
&- \gamma(u_k - K^j x_k)^\top B^\top \hat{P}^{j+1}(Be_{k1} + De_{k2}) \\
&- \gamma e_{k1}^\top B^\top \hat{P}^{j+1} x_{k+1} \\
&- \gamma e_{k1}^\top B^\top \hat{P}^{j+1}(Be_{k1} + De_{k2}) \\
&- \gamma(u_k - K^j x_k)^\top B^\top \hat{P}^{j+1} A_k x_k \\
&- \gamma e_{k1}^\top B^\top \hat{P}^{j+1} A_k x_k \\
&- \gamma(v_k - L^j x_k)^\top D^\top \hat{P}^{j+1} x_{k+1} \\
&- \gamma(v_k - L^j x_k)^\top B^\top \hat{P}^{j+1}(Be_{k1} + De_{k2}) \\
&- \gamma e_{k2}^\top D^\top \hat{P}^{j+1} x_{k+1} \\
&- \gamma e_{k2}^\top D^\top \hat{P}^{j+1}(Be_{k1} + De_{k2}) \\
&- \gamma(v_k - L^j x_k)^\top D^\top \hat{P}^{j+1} A_k x_k \\
&- \gamma e_{k2}^\top D^\top \hat{P}^{j+1} A_k x_k.
\end{aligned} \tag{49}$$

Substituting (32) into (49) and then eliminating the common terms yields

$$\begin{aligned}
&x_{k+1}^\top \hat{P}^{j+1}(Be_{k1} + De_{k2}) \\
&= (u_k - K^j x_k)^\top B^\top \hat{P}^{j+1}(Be_{k1} + De_{k2}) \\
&+ x_k^\top A_k^\top \hat{P}^{j+1}(Be_{k1} + De_{k2}) \\
&+ (v_k - L^j x_k)^\top D^\top \hat{P}^{j+1}(Be_{k1} + De_{k2}).
\end{aligned} \tag{50}$$

Substituting (50) into (49) yields

$$\begin{aligned}
&x_k^\top \hat{P}^{j+1} x_k - \gamma x_{k+1}^\top \hat{P}^{j+1} x_{k+1}\\
&= x_k^\top \bar{Q} x_k + x_k^\top (K^j)^\top \hat{P}^{j+1} K^j x_k\\
&\quad + x_k^\top (L^j)^\top \hat{P}^{j+1} L^j x_k\\
&\quad - \gamma (u_k - K^j x_k)^\top B^\top \hat{P}^{j+1} x_{k+1}\\
&\quad - \gamma (u_k - K^j x_k)^\top B^\top \hat{P}^{j+1} A_k x_k\\
&\quad - \gamma (v_k - L^j x_k)^\top D^\top \hat{P}^{j+1} x_{k+1}\\
&\quad - \gamma (v_k - L^j x_k)^\top D^\top \hat{P}^{j+1} A_k x_k. \qquad (51)
\end{aligned}$$

Comparing (32) with (51) shows that $\hat{P}^{j+1}$ is equal to P^{j+1}. This completes the proof. ■

Remark 2. Theorem 1 indicates that the addition of the probing noise e_{k1} and e_{k2} did not bias the solution of P^{j+1} in off-policy BE (32). Through (33) and (34), it can be concluded that the iterative control policies u_k^{j+1} and v_k^{j+1} are only affected by P^{j+1}. Therefore, the solutions of the iterative control policies u_k^{j+1} and v_k^{j+1} are also unbiased. Consequently, this off-policy RL method avoids the bias in the solution caused by the addition of probing noise.

Remark 3. The solution of LS equation (40) is equivalent to the solution of off-policy BE (32). Likewise, the solutions of the iterative control policies in (41) and (42) are equivalent to the solutions of the iterative control policies in (33) and (34). Therefore, according to the proof of theorem 1, the optimal control policies solved by Algorithm 1 are unbiased.

IV. Simulation

In this section, the proposed model-free off-policy RL algorithm 1 is used to control the rotating inverted pendulum model mentioned in [30]. Three cases demonstrate the impact of the discount factor and probing noise. In Cases 1 and 2, the magnitudes of the probing noise are varied while the frequencies are held constant. In Case 3, the frequencies of the probing noise are varied. Consider the discrete-time rotating inverted pendulum model used in [30]:

$$x_{k+1} = [A + \Delta A(p)]x_k + Bu_k \qquad (52)$$

where system matrices

$$A = \begin{bmatrix} 1.0008 & 0.005 & 0 & 0 \\ 0.3164 & 1.008 & 0 & 0 \\ -0.0004 & 0 & 1 & 0.005 \\ -0.1666 & -0.0004 & 0 & 1 \end{bmatrix}$$

$$B = \begin{bmatrix} -0.0065 & -2.6043 & 0.0101 & 4.0210 \end{bmatrix}^\top.$$

The uncertainty in the system is defined as

$$\Delta A(p) = S\phi_A(p) \qquad (53)$$

where

$$S = \begin{bmatrix} 0.0064 & -2.5648 & 0.0101 & 4.0210 \end{bmatrix}^\top$$
$$\phi_A(p) = p \times sin(0.6k) \begin{bmatrix} 0.21 & 0.1 & 0.04 & 0.03 \end{bmatrix}.$$

and $p \in \pm 1.1$. For the purpose of simulation, the value of p is set to -0.7, satisfying condition (5). The parameter F associated with uncertainty in (5) is selected as

$$F = \begin{bmatrix} 48.4 & 24.2 & 9.68 & 7.26 \\ 24.2 & 12.1 & 4.84 & 3.63 \\ 9.68 & 4.84 & 1.93 & 1.45 \\ 7.26 & 3.63 & 1.45 & 1.08 \end{bmatrix}$$

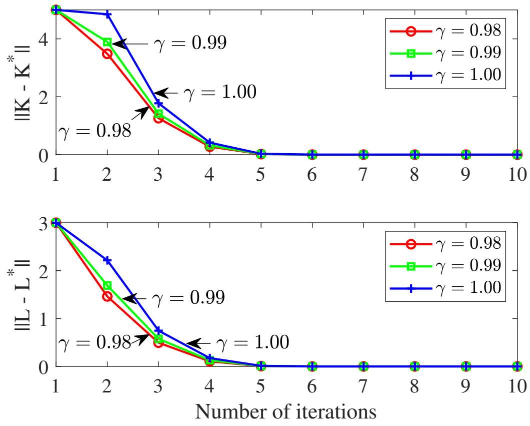

Fig. 1. Case 1: Convergence of K and L in off-policy RL .

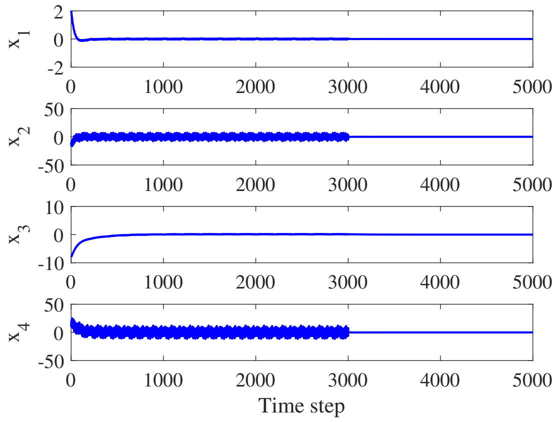

Fig. 2. Case 1: Auxiliary system state in off-policy RL.

and $\varepsilon = 0.001$. The parameters α in auxiliary system (7) is designed as $\alpha = 0.02$. The weighting matrices in cost function (8) are designed as $Q = I_4$, $R_1 = 3$, $R_2 = 4$ and $\beta = 1$. The predetermined error bounds are set as $\delta_1 = \delta_2 = 0.0001$. The initial state of the system is set to $x_0 = \begin{bmatrix} 2 & -10 & 8 & 10 \end{bmatrix}^\top$. The exact solution of P^* with $\gamma = 1$ is

$$P^* = 10^4 \times \begin{bmatrix} 1.6980 & 0.2437 & 0.1305 & 0.1532 \\ 0.2437 & 0.0445 & 0.0266 & 0.0277 \\ 0.1305 & 0.0266 & 0.0577 & 0.0169 \\ 0.1532 & 0.0277 & 0.0169 & 0.0176 \end{bmatrix}$$

and the optimal feedback gains with $\gamma = 1$ are

$$K^* = \begin{bmatrix} 3.7579 & 0.7596 & 0.2207 & 0.2437 \end{bmatrix}$$
$$L^* = -\begin{bmatrix} 1.2667 & 0.1885 & 0.1223 & 0.1193 \end{bmatrix}.$$

First, the off-policy RL algorithm will be used to solve the optimal feedback gains. The admissible behavior policies of this algorithm are designed as

$$K = \begin{bmatrix} 4.1705 & 0.7643 & 0.1679 & 0.2305 \end{bmatrix}$$
$$L = \begin{bmatrix} 15.6834 & 2.4478 & 1.5514 & 1.5536 \end{bmatrix}.$$

200 data samples will be collected at each iteration to solve LS equation (40). The following simulation examples will be provided. During the learning process from Case 1 to Case 3, probing noise persists until the 3000th step.

Case 1 : The probing noise is considered as

$$e_{k1} = cos(0.5k) + cos(2k) + cos(10k)$$
$$e_{k2} = sin(1.7k) + sin(k).$$

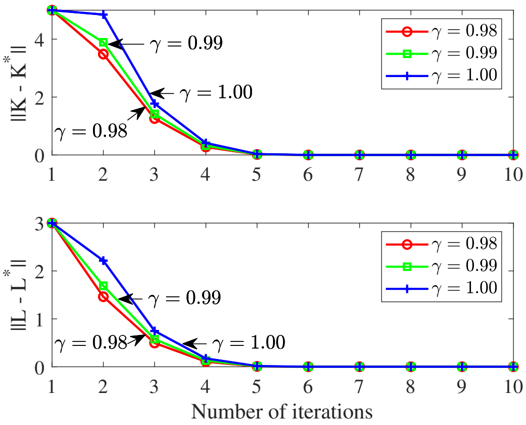

Fig. 3. Case 2: Convergence of K and L in off-policy RL.

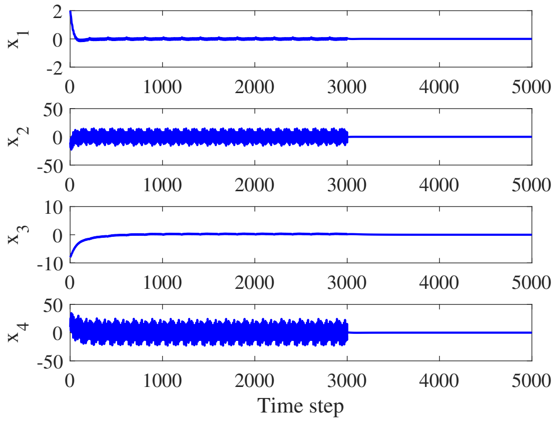

Fig. 4. Case 2: Auxiliary system state in off-policy RL.

Fig. 1 shows the norms of the learning error of iterative feedback gains $(K^j\ and\ L^j)$ and optimal feedback gains $(K^*\ and\ L^*)$ with different discount factors during the learning process when the probing noise is Case 1. When the discount factors are 0.98, 0.99, and 1, respectively, the number of iterations in the learning process is 5, 6, and 9, respectively. Fig. 2 shows the state of auxiliary system (7) with the admissible control policies applied during the learning process. It can be seen that the iterative control policies make all system states go to 0.

Case 2 : The probing noise is considered as

$$e_{k1} = 2cos(0.5k) + 2cos(2k) + 2cos(10k)$$
$$e_{k2} = 2sin(1.7k) + 2sin(k).$$

Fig. 3 shows the norms of the learning error of iterative feedback gains $(K^j\ and\ L^j)$ and optimal feedback gains $(K^*\ and\ L^*)$ with different discount factors during the learning process when the probing noise is Case 2. When the discount factors are 0.98, 0.99, and 1, respectively, the number of iterations in the learning process is 5, 9, and 13, respectively. Fig. 4 shows the state of auxiliary system (7) with the admissible control policies applied during the learning process. It can be seen that the iterative control policies make all system states go to 0.

Case 3 : The probing noise is considered as

$$e_{k1} = cos(1.5k) + cos(k) + cos(2k)$$
$$e_{k2} = sin(3k) + sin(2k).$$

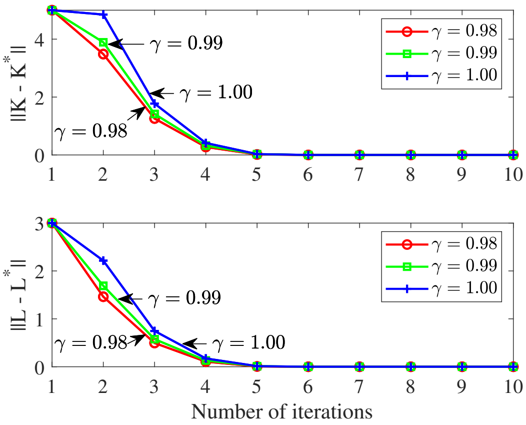

Fig. 5. Case 3: Convergence of K and L in off-policy RL .

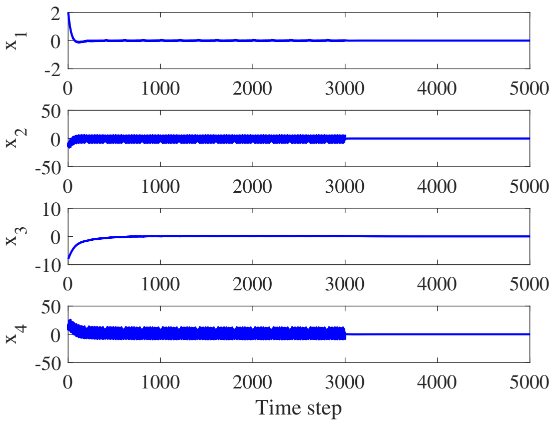

Fig. 6. Case 3: Auxiliary system state in off-policy RL.

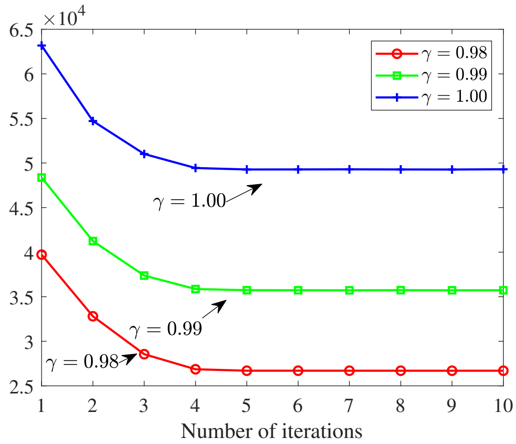

Fig. 7. The value function with different discount factors.

Fig. 5 shows the norms of the learning error of iterative feedback gains $(K^j\ and\ L^j)$ and optimal feedback gains $(K^*\ and\ L^*)$ with different discount factors during the learning process when the probing noise is Case 3. When the discount factors are 0.98, 0.99, and 1, respectively, the number of iterations in the learning process is 5, 7, and 16, respectively. Fig. 6 shows the state of auxiliary system (7) with the admissible control policies applied during the learning process. It can be seen that the iterative control policies make all system states go to 0.

In the next, the approximate optimal feedback control gain K^* obtained using the model-free off-policy algorithm 1 with different discount factors is used to solve the robust

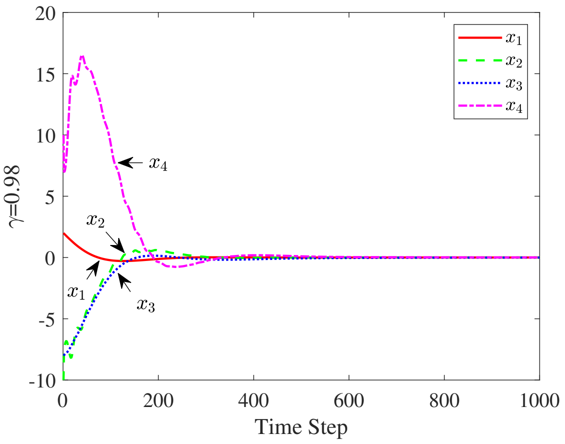

Fig. 8. Uncertian system state with γ = 0.98.

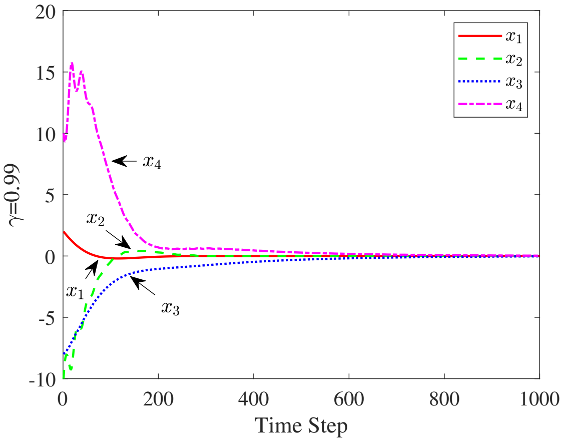

Fig. 9. Uncertian system state with γ = 0.99.

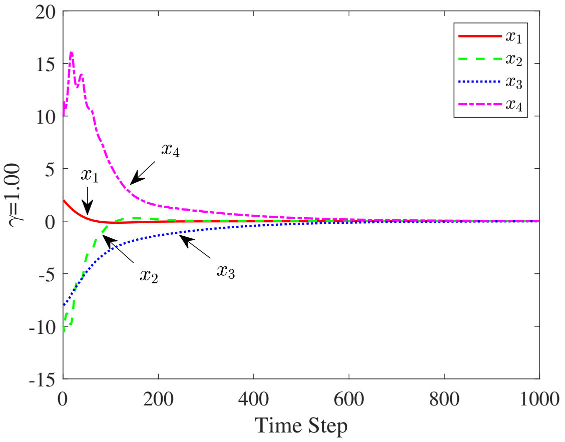

Fig. 10. Uncertian system state with γ = 1.

control problem of uncertain system (52). Fig. 7 shows the convergence process of value functions during the learning process with different discount factors. Figs. 8–10 show the system state trajectories with different discount factors.

V. Conclusion

This paper proposed a model-free off-policy RL algorithm incorporating a discount factor to solve robust control issure of uncertain DLSs using only measured data. By building an auxiliary system, the robust control issure was transformed into an optimal control issure and then solved using the proposed algorithm. Theoretically, it was proved that this algorithm did not produce biased results after adding probing noise. Additionally, this algorithm converged faster and incurred lower control costs. Finally, a simulation example verified the effectiveness of the proposed algorithm.

References

[1] M. Benallouch, G. Schutz, D. Fiorelli, and M. Boutayeb, “$\mathcal{H}_\infty$ model predictive control for discrete-time switched linear systems with application to drinking water supply network,” *Journal of Process Control*, vol. 24, no. 6, pp. 924–938, 2014.

[2] S. Solmaz, R. Shorten, K. Wulff, and F. O. Cairbre, “A design methodology for switched discrete time linear systems with applications to automotive roll dynamics control,” *Automatica*, vol. 44, no. 9, pp. 2358–2363, 2008.

[3] N. Meskin and K. Khorasani, “Fault detection and isolation of discrete-time Markovian jump linear systems with application to a network of multi-agent systems having imperfect communication channels,” *Automatica*, vol. 45, no. 9, pp. 2032–2040, 2009.

[4] L. He, X. Zhang, X. Wang, and J. Zhou, “Performance-guaranteed Control for Discrete-time Systems Under Communication Constraints: An Event-triggered Mechanism and Quantized Data-based Protocol,” *IAENG International Journal of Computer Science*, vol. 51, no. 10, pp. 1570–1578, 2024.

[5] S. Santra, M. Joby, M. Sathishkumar, and S. M. Anthoni, “LMI approach-based sampled-data control for uncertain systems with actuator saturation: Application to multi-machine power system,” *Nonlinear Dynamics*, vol. 107, pp. 967–982, 2022.

[6] N. Setiawan, G. N. Putu Pratama, H. A. Winarno, S. Herdjunanto, and A. I. Cahyadi, “Attitude tracking control on SO(3) group with linearization on moving operating point for transporting quadrotor,” *IAENG International Journal of Computer Science*, vol. 48, no. 4, pp. 940–951, 2021.

[7] Y. Wang, “Exponential stabilization and $\mathcal{L}_2$-gain for a class of nonlinear switched uncertain systems with mixed delays,” *Engineering Letters*, vol. 27, no. 3, pp. 458–466, 2019.

[8] C. Aguiar, D. Leite, D. Pereira, G. Andonovski, and I. Škrjanc, “Nonlinear modeling and robust LMI fuzzy control of overhead crane systems,” *Journal of the Franklin Institute*, vol. 358, no. 2, pp. 1376–1402, 2021.

[9] N. Gunasekaran, N. M. Thoiyab, Q. Zhu, J. Cao, and P. Muruganantham, “New global asymptotic robust stability of dynamical delayed neural networks via intervalized interconnection matrices,” *IEEE Transactions on Cybernetics*, vol. 52, no. 11, pp. 11 794–11 804, 2022.

[10] J. Zhou, J. H. Park, and Q. Kong, “Robust resilient $\mathcal{L}_2 - \mathcal{L}_\infty$ control for uncertain stochastic systems with multiple time delays via dynamic output feedback,” *Journal of the Franklin Institute*, vol. 353, no. 13, pp. 3078–3103, 2016.

[11] C. E. De Souza and D. Coutinho, “Robust stability and control of uncertain linear discrete-time periodic systems with time-delay,” *Automatica*, vol. 50, no. 2, pp. 431–441, 2014.

[12] C. F. Morais, M. F. Braga, R. C. Oliveira, and P. L. Peres, “Reduced-order dynamic output feedback control of uncertain discrete-time Markov jump linear systems,” *International Journal of Control*, vol. 90, no. 11, pp. 2368–2383, 2017.

[13] C. M. Massera, M. H. Terra, and D. F. Wolf, “Optimal guaranteed cost control of discrete-time linear systems subject to structured uncertainties,” *International Journal of Control*, vol. 94, no. 4, pp. 1132–1142, 2021.

[14] Y. Jiang, L. Liu, and G. Feng, “Fully distributed adaptive control for output consensus of uncertain discrete-time linear multi-agent systems,” *Automatica*, vol. 162, p. 111531, 2024.

[15] X.-H. Liu, Z. Xue, J. Pang, S. Jiang, F. Xu, and Y. Yu, “Regret minimization experience replay in off-policy reinforcement learning,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 604–17 615, 2021.

[16] Y. Yang, Z. Guo, H. Xiong, D.-W. Ding, Y. Yin, and D. C. Wunsch, “Data-driven robust control of discrete-time uncertain linear systems via off-policy reinforcement learning,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 12, pp. 3735–3747, 2019.

[17] F. Lin, *Robust Control Design: An Optimal Control Approach*. Chichester, U.K.: Wiley, 2007.

[18] R. A. Horn and C. R. Johnson, *Matrix Analysis*, New York, NY, USA: Cambridge Univ. Press, 2012.

[19] Z. Yan, X. Huang, and J. Liang, “Aperiodic sampled-data control for stabilization of memristive neural networks with actuator saturation: A dynamic partitioning method,” *IEEE Transactions on Cybernetics*, vol. 53, no. 3, pp. 1725–1737, 2023.

[20] L. A. Rios-Norena, J. S. Velez-Ramirez, and E. Giraldo, “Real-time optimal embedded control of a double inverted pendulum,” *IAENG International Journal of Computer Science*, vol. 49, no. 2, pp. 341–348, 2022.

[21] J. Zhou, J. Dong, and S. Xu, "Asynchronous dissipative control of discrete-time fuzzy Markov jump systems with dynamic state and input quantization," *IEEE Transactions on Fuzzy Systems*, vol. 31, no. 11, pp. 3906–3920, 2023.

[22] X. Wang, X. Qin, Y. Ji, T. Jiang, and J. Zhou, "Mean-square asymptotic synchronization of complex dynamical networks subject to communication delay and switching topology," *Physica Scripta*, vol. 98, no. 10, p. 105214, 2023.

[23] L. He, W. Wu, G. Yao, and J. Zhou, "Input-to-state stabilization of delayed semi-Markovian jump neural networks via sampled-data control," *Neural Processing Letters*, vol. 55, no. 3, pp. 3245–3266, 2023.

[24] L. Yao, Z. Wang, X. Huang, Y. Li, Q. Ma, and H. Shen, "Stochastic sampled-data exponential synchronization of Markovian jump neural networks with time-varying delays," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 2, pp. 909–920, 2021.

[25] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal Control*. Hoboken, NJ, USA: Wiley, 2012.

[26] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Belmont, MA, USA: Athena Scientific, 1996.

[27] D. Wang, J. Ren, M. Ha, and J. Qiao, "System stability of learning-based linear optimal control with general discounted value iteration," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 9, pp. 6504–6514, 2023.

[28] G. Tao, *Adaptive Control Design and Analysis*. Hoboken, NJ, USA: Wiley, 2003.

[29] G. B. D. Sousa and P. H. M. Rêgo, "Convergence and numerical stability of action-dependent heuristic dynamic programming algorithms based on RLS learning for online DLQR optimal control," *International Journal of Computational Science and Engineering*, vol. 20, no. 3, pp. 317–334, 2019.

[30] N. S. Tripathy, I. Kar, and K. Paul, "Stabilization of uncertain discrete-time linear system with limited communication," *IEEE Transactions on Automatic Control*, vol. 62, no. 9, pp. 4727–4733, 2016.