Data-Driven Robust Control of Discrete-Time Uncertain Linear Systems via Off-policy Reinforcement Learning

Yongliang Yang, *Member, IEEE,* Zhishan Guo, *Member, IEEE,* Haoyi Xiong, *Member, IEEE,* Da-Wei Ding, *Member, IEEE,* Yixin Yin, *Member, IEEE,* and Donald C. Wunsch II, *Fellow, IEEE*

***Abstract*—This paper presents a model-free solution to the robust stabilization problem of discrete-time linear dynamical systems with bounded and mismatched uncertainty. An optimal controller design method is derived to solve the robust control problem, which results in solving an algebraic Riccati Equation (ARE). It is shown that the optimal controller obtained by solving the ARE can robustly stabilize the uncertain system. To develop a model-free solution to the translated ARE, off-policy reinforcement learning (RL) is employed to solve the problem in hand without the requirement of system dynamics. In addition, the comparisons between on- and off-policy RL methods are presented regarding the robustness to probing noise and the dependency on system dynamics. Finally, a simulation example is carried out to validate the efficacy of the presented off-policy RL approach.**

***Index Terms*—system uncertainty, robust control, reinforcement learning, on-policy, off-policy, model free.**

I. Introduction

Robust control of uncertain dynamical systems has received considerable attention in the control community as well as many other fields such as chemical process, power systems, robotics, and aerospace engineering. It is of paramount importance to achieve robust performance and/or stability in the presence of bounded system uncertainties, such as external disturbances, unmodeled dynamics, and time-varying system parameters, and so on [1]–[5]. There has been extensive research on the robust control theory, including the frequency-domain analysis [6], [7], optimization methods [8] and time-domain method [9]. However, in most existing results, the system dynamics is required for the robust controller design, which might be vulnerable to exhaustive modeling and potential attacks. The main concern of this paper is to obviate the requirement of complete knowledge of system dynamics for the robust stabilization problem of discrete-time linear systems with mismatched uncertainty.

Recently, the relation between robust stabilization and optimal controller design has been studied in [10], in which it is shown that the optimal controller of an auxiliary system can stabilize the uncertain system. Solving the optimal control problem results in solving the algebraic Riccati equation (ARE) for linear systems or Hamilton-Jacobi-Bellman (HJB) equation for nonlinear systems. However, for general nonlinear systems, the HJB equation is essentially a nonlinear partial differential equation, of which the analytical solution might not exist. Besides, dynamic programming has to be implemented backward-in-time which often makes the computation unavailable with increasing dimension [11]. Therefore, approximate dynamic programming (ADP) algorithms [12] are developed to approximately solve the HJB equation forward-in-time by using function approximation techniques, such as neural networks [13]. Variants of ADP methods are developed since then [14]–[16], including iterative off-line ADP [17], [18] and model-based on-line ADP [19] and identification-based ADP [20]–[22]. Even though the identification-based ADP does not require the system dynamics, the accuracy of the system identification has an impact on the control performance. Therefore, the data-driven controller design of which the performance does not depend on the complete knowledge of system dynamics is desired.

Reinforcement learning (RL) techniques have been successfully applied to solve the decision-making problems when the agent is interacting with an uncertain environment [23], [24]. In general, RL approaches can be divided into on- and off-policy RL methods. In the on-policy RL method, it is required that the control policy to be evaluated has to be applied to the systems. Typical on-policy RL method is

Manuscript submitted January 31, 2019. This work was supported in part by the National Natural Science Foundation of China under grant No. 61333002, No. 61473032 and No. 61873028, in part by the Fundamental Research Funds for the China Central Universities of USTB under grant No. FRF-TP-18-031A1 and No. FRF-BD-17-002A, in part by the China Post-Doctoral Science Foundation under Grant 2018M641198, in part by the Mary K. Finley Endowment, in part by the Missouri S&T Intelligent Systems Center and in part by the Army Research Laboratory under Cooperative Agreement Number W911NF-18-2-0260. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Yongliang Yang, Dawei Ding and Yixin Yin are with the School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China, and also with the Key Laboratory of Knowledge Automation for Industrial Process, Ministry of Education, Beijing 100083, China. (e-mail: yangyongliang@ieee.org; dingdawei@ustb.edu.cn; yyx@ies.ustb.edu.cn).

Zhishan Guo is with the Department of Electric and Computer Engineering, University of Central Florida, Orlando, Florida, USA (e-mail: Zhishan.Guo@ucf.edu).

Haoyi Xiong is with the Big Data Laboratory, Baidu Research, Beijing 100193, China, and also with the National Engineering Laboratory for Deep Learning Technology and Applications, Baidu Inc, Beijing 100193, China (e-mail: haoyi.xiong.fr@ieee.org).

Donald C. Wunsch II is with the Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla, MO 65401, USA (e-mail: dwunsch@mst.edu).

SARSA algorithm [25], which updates the value function based on the experience obtained by executing some policy. In contrast, the off-policy RL approach aims at learning the optimal control policy when another admissible policy, not necessarily optimal, is interacting with the environment. The well-known Q-learning algorithm belongs to the off-policy RL class because the learning policy is different from the policy to be carried out [26]. Off-policy RL has been applied to deal with optimal regulation problems [27], optimal tracking problem [28] and differential games [29], [30] for single-agent systems. Recently, off-policy RL has also been applied to output synchronization problem [31]–[33], containment control problem [34] and graphical games [35] for multi-agent systems. To the authors' knowledge, the off-policy RL method has not been applied to the robust stabilization problem of discrete-time uncertain systems yet. This paper develops an off-policy RL method to obtain the robust controller of the uncertain system without requiring the system dynamics a priori.

Motivated by the above-mentioned work, in this paper, an off-policy RL-based method is developed for the robust controller design of discrete-time linear systems in the presence of mismatched uncertainty. First, the robust control problem for the original uncertain system is translated to the optimal control problem for an auxiliary system with a properly modified reward function. Then, the sufficient condition that guarantees the translation equivalence, i.e., the optimal control for the auxiliary system can robustly stabilize the original uncertain system, is also discussed. Meanwhile, on- and off-policy RL methods are compared and discussed in detail. The main contributions of this paper are as follows.

1) Model-free robust controller design is derived to achieve robust stabilization of the discrete-time uncertain system by solving an optimal control problem for an auxiliary system with a modified performance function. Sufficient condition that guarantees the optimal controller could ensure robust stabilization of the discrete-time uncertain systems is also provided.
2) Variants of on- and off-policy RL method are derived. In addition, comparisons between on- and off-policy RL methods are discussed in terms of the robustness to the probing noise and the dependency on the system dynamics.

The remainder of this paper is organized as follows. Section II describes the robust control problem of the discrete-time linear system with mismatched uncertainty. In Section III, the robust control problem is translated into the optimal control problem of an auxiliary system. The sufficient condition, which guarantees the optimal control policy of the auxiliary system can robustly stabilize the uncertain system, is also given in Section III. Two types of RL methods, on- and off-policy RL, are discussed in detail in Section IV and V, respectively. In Section VI, a simulation is conducted to demonstrate the validity of the proposed approach. Finally, concluding remarks and future works are presented in Section VII.

II. Problem Formulation

In this paper, a class of discrete-time nonlinear systems with uncertainty is considered, which can be described as

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

with the system state $x_k \in \mathbb{R}^n$, the control input $u_k \in \mathbb{R}^m$, the drift dynamics $A+\Delta \in \mathbb{R}^{n\times n}$, the input dynamics $B \in \mathbb{R}^{n\times m}$ and p is a vector of uncertain parameters which is restricted to a prescribed bounded and compact set Ω. The drift dynamics $A+\Delta \in \mathbb{R}^{n\times n}$ consists of nominal part A and uncertain part $\Delta \in \mathbb{R}^{n\times n}$. In addition, the nominal system of (1) is

$$x_{k+1} = A x_k + B u_k. \tag{2}$$

Moreover, the nominal system (2) satisfies the following assumption.

Assumption 1. *The pair (A, B) is stabilizable.* □

The system uncertainty Δ can be classified as matched and mismatched uncertainty according to its relation to the input dynamics B [10]. To be specific, the uncertainty in system (1) belongs to the type of matched uncertainty if $\Delta(p)$ can be expressed as

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

That is, the system uncertainty Δ is in the space spanned by the columns of input matrix B. For the case of mismatched uncertainty, Δ cannot be expressed in the form of (3). Moreover, mismatch uncertainty can be decomposed of match part and mismatched part

$$\begin{aligned}\Delta(p) &= S\phi(p) \\ &= BB^{\dagger}S\phi(p) + \left(I - BB^{\dagger}\right)S\phi(p), \forall p \in \Omega,\end{aligned} \tag{4}$$

where $B^{\dagger}$ is the pseudo-inverse of B, $S \neq B$, $S \in \mathbb{R}^{n\times r}$ is a known weight matrix and $\phi(p) \in \mathbb{R}^{r\times n}$ is unknown perturbation. In this paper, the perturbation $\phi(p)$ is bounded in the following sense.

Assumption 2. *There exists a positive semi-definite matrix F such that*

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

where ε is a positive constant. □

The robust control problem of system (1) of interest in this paper can be formulated as follows.

Problem 1. (Robust Control Problem) *Find a state feedback control law $u_k = Kx_k$ such that the close-loop system*

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

is asymptotically stable for $\forall p \in \Omega$. □

For the purpose of designing the robust control $u_k = Kx_k$ to stabilize system (1), the state feedback gain K is designed by ARE approach in optimal control theory. By introducing an extra term Dv_k to the nominal system (2) one can obtain an auxiliary system as,

$$x_{k+1} = A x_k + B u_k + D v_k, \tag{7}$$

where α is a positive constant, $D = \alpha\left(I - BB^{\dagger}\right)S \in \mathbb{R}^{n\times r}$ and r is the rank of B. Then, the optimal control problem of the auxiliary system (7), which is closely related to the above robust control problem, can be described as follows.

Problem 2. (Optimal Control Problem) *Find state feedback control laws $u_k = K^* x_k$ and $v_k = L^* x_k$ such that the performance*

$$V(x_k) = \frac{1}{2}\sum_{j=k}^{\infty}\left(x_j^{\mathrm{T}} Q x_j + x_j^{\mathrm{T}} F x_j + \beta^2 x_j^{\mathrm{T}} x_j + u_j^{\mathrm{T}} R_1 u_j + v_j^{\mathrm{T}} R_2 v_j\right) \tag{8}$$

with respect to the auxiliary system (7) is minimized, where $Q \geq 0$ is a positive semi-definite matrix, $R_1 \succ 0$ and $R_2 \succ 0$ are positive definite matrices and β is a positive constant. □

For simplicity, the terms in the summation in (8) is denoted as

$$r(x_k, u_k, v_k) = x_k^{\mathrm{T}} Q x_k + x_k^{\mathrm{T}} F x_k + \beta^2 x_k^{\mathrm{T}} x_k + u_k^{\mathrm{T}} R_1 u_k + v_k^{\mathrm{T}} R_2 v_k.$$

which is referred to as the utility function.

Remark 1. *As shown later in Section III, under some specific conditions, K^* can be used as robust state feedback gain to stabilize the uncertain system (1). That is, the robust control problem of uncertain system (1) can be translated to the optimal control problem of the auxiliary system (7) with the performance defined in (8). Note that the control input v_k only appears in the auxiliary system (7). Therefore, the feedback gain L^* does not affect the system (1) directly and v_k is referred to as virtual control.* □

III. Robust Controller Design Using ARE Approach

In this section, ARE approach in optimal control theory is used to solve the robust control problem of the uncertain system (1). The robust control problem of the uncertain system (1) is transformed to an optimal control problem of the auxiliary system (7) with respect to the performance (8). The condition that guarantees the equivalence between the robust control problem and the optimal control problem is provided.

To begin with, the following are required for the subsequent discussions.

Definition 1. (**Admissible Control**) *For the auxiliary system (7), the control mappings $u(x)$ and $v(x)$ are said to be admissible with respect to performance (8) if*

- *$u(x_k)$ and $v(x_k)$ are continuous;*
- *$u(0) = v(0) = 0$;*
- *$u(x_k)$ and $v(x_k)$ stabilize the auxiliary system (7);*
- *The value function $V(x_k)$ w.r.t. the policies $u(\cdot)$ and $v(\cdot)$ is finite for $\forall x_k$.* □

Lemma 1. *For arbitrary admissible control $u(x_k) = Kx_k$ and $v(x_k) = Lx_k$, the performance function $V(x_k)$ in (8) is quadratic in x_k, for $\forall x_k \in \mathbb{R}^n$.* □

Proof: Taking the control $u(x_k) = Kx_k$ and $v(x_k) = Lx_k$ into the auxiliary system (7), the closed-loop dynamics should be

$$x_{k+1} = (A + BK + DL)x_k = \bar{A}x_k.$$

Therefore, $x_{k+j} = \bar{A}^j x_k,\ \forall j = 0, 1, 2, \cdots$. Now inserting $u(x_k) = Kx_k$ and $v(x_k) = Lx_k$ into the reward function $r(x_k, u_k, v_k)$ yields

$$\begin{aligned} r(x_k, u_k, v_k) &= x_k^{\mathrm{T}}\left(Q + F + \beta^2 I + K^{\mathrm{T}} R_1 K + + L^{\mathrm{T}} R_2 L\right)x_k \\ &= x_k^{\mathrm{T}} \mathcal{Q} x_k, \end{aligned}$$

where $\mathcal{Q} = Q + F + \beta^2 I + K^{\mathrm{T}} R_1 K + L^{\mathrm{T}} R_2 L$. Therefore, the value function is equivalent to

$$\begin{aligned} V(x_k) &= \sum_{j=k}^{\infty} r(x_j, u_j, v_j) = \sum_{j=k}^{\infty} x_j^{\mathrm{T}} \mathcal{Q} x_j \\ &= x_k^{\mathrm{T}}\left(\sum_{j=k}^{\infty}\left(\bar{A}^{\mathrm{T}}\right)^{j-k}\mathcal{Q}\bar{A}^{j-k}\right)x_k. \end{aligned}$$

This completes the proof. ■

To solve the optimal control problem of the auxiliary system (7) with the performance (8), the optimal control laws $u_k = K^* x_k$ and $v_k = L^* x_k$ are derived in the following theorem.

Theorem 1. *Suppose that there exists a positive definite solution $P \succ 0$ of the following algebraic Riccati equation (ARE)*

$$\begin{aligned} 0 = &-\begin{bmatrix} B^{\mathrm{T}}PA \\ D^{\mathrm{T}}PA \end{bmatrix}^{\mathrm{T}}\begin{bmatrix} R_1 + B^{\mathrm{T}}PB & B^{\mathrm{T}}PD \\ D^{\mathrm{T}}PB & R_2 + D^{\mathrm{T}}PD \end{bmatrix}^{-1} \\ &\begin{bmatrix} B^{\mathrm{T}}PA \\ D^{\mathrm{T}}PA \end{bmatrix} + A^{\mathrm{T}}PA - P + \bar{Q}, \end{aligned} \tag{9}$$

where $\bar{Q} = Q + F + \beta^2 I$. Then the optimal control of system (7) with respect to the performance function (8) can be expressed as $u_k^* = K^* x_k$ and $v_k^* = L^* x_k$ with gains K^* and L^* satisfying

$$\begin{aligned} K^* = &-\left[R_1 + B^{\mathrm{T}}PB - B^{\mathrm{T}}PD\left(R_2 + D^{\mathrm{T}}PD\right)^{-1}D^{\mathrm{T}}PB\right]^{-1} \\ &\left[B^{\mathrm{T}}PA - B^{\mathrm{T}}PD\left(R_2 + D^{\mathrm{T}}PD\right)^{-1}D^{\mathrm{T}}PA\right], \end{aligned} \tag{10}$$

$$\begin{aligned} L^* = &-\left[R_2 + D^{\mathrm{T}}PD - D^{\mathrm{T}}PB\left(R_1 + B^{\mathrm{T}}PB\right)^{-1}B^{\mathrm{T}}PD\right]^{-1} \\ &\left[D^{\mathrm{T}}PA - D^{\mathrm{T}}PB\left(R_1 + B^{\mathrm{T}}PB\right)^{-1}B^{\mathrm{T}}PA\right]. \end{aligned} \tag{11}$$

□

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

$$V(x_k) = V(x_{k+1}) + r(x_k, u_k, v_k). \tag{12}$$

Define the Hamiltonian as

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

From Lemma 1, the value function in (8) can be denoted as

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

Based on [36], the necessary conditions for optimal control u_k^* and v_k^* is given by

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

Considering the Hamiltonian and the quadratic value function, (14) is equivalent to:

$$\begin{aligned}&\left[\begin{array}{cc}\left(R_1+B^{\mathrm{T}}PB\right) & B^{\mathrm{T}}PD\\ D^{\mathrm{T}}PB & \left(R_2+D^{\mathrm{T}}PD\right)\end{array}\right]\left[\begin{array}{c}u_k^*\\ v_k^*\end{array}\right]\\ &=-\left[\begin{array}{c}B^{\mathrm{T}}PA\\ D^{\mathrm{T}}PA\end{array}\right]x_k.\end{aligned}$$

Denote

$$\begin{aligned}\mathcal{E}&=B^{\mathrm{T}}PA,\\ \mathcal{G}&=D^{\mathrm{T}}PA,\\ \mathcal{M}&=\left[\begin{array}{cc}\mathcal{M}_{11} & \mathcal{M}_{12}\\ \mathcal{M}_{21} & \mathcal{M}_{22}\end{array}\right]\\ &=\left[\begin{array}{cc}\left(R_1+B^{\mathrm{T}}PB\right) & B^{\mathrm{T}}PD\\ D^{\mathrm{T}}PB & \left(R_2+D^{\mathrm{T}}PD\right)\end{array}\right].\end{aligned}$$

then the optimal control u_k^* and v_k^* can be expressed as

$$\left[\begin{array}{c}u_k^*\\ v_k^*\end{array}\right]=-\mathcal{M}^{-1}\left[\begin{array}{c}\mathcal{E}\\ \mathcal{G}\end{array}\right]x_k.$$

Let $\mathcal{N}=\mathcal{M}^{-1}$ be partitioned into the block form as $\mathcal{N}=\left[\begin{array}{cc}\mathcal{N}_{11} & \mathcal{N}_{12}\\ \mathcal{N}_{21} & \mathcal{N}_{22}\end{array}\right]$. Based on the matrix inversion lemma [37], $\mathcal{N}$ can be expressed as:

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

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

$$K^*=-\left(\mathcal{N}_{11}\mathcal{E}+\mathcal{N}_{12}\mathcal{G}\right), \tag{15}$$

$$L^*=-\left(\mathcal{N}_{21}\mathcal{E}+\mathcal{N}_{22}\mathcal{G}\right). \tag{16}$$

By collecting above results, (15) and (16) are equivalent to (10) and (11).

Let $\bar{Q}=Q+F+\beta^2I$. The optimal control u_k^* and v_k^* satisfy

$$\begin{aligned}0&=\min_{u_k,v_k}H\left(x_k,u_k,v_k\right)=H\left(x_k,u_k^*,v_k^*\right)\\ &=\left[\begin{array}{c}u_k^*\\ v_k^*\end{array}\right]^{\mathrm{T}}\left[\begin{array}{cc}R_1+B^{\mathrm{T}}PB & B^{\mathrm{T}}PD\\ D^{\mathrm{T}}PB & R_2+D^{\mathrm{T}}PD\end{array}\right]\left[\begin{array}{c}u_k^*\\ v_k^*\end{array}\right]\\ &+\left[\begin{array}{c}u_k^*\\ v_k^*\end{array}\right]^{\mathrm{T}}\left[\begin{array}{c}B^{\mathrm{T}}PA\\ D^{\mathrm{T}}PA\end{array}\right]x_k+x_k^{\mathrm{T}}\left(A^{\mathrm{T}}PA-P\right)x_k\\ &+x_k^{\mathrm{T}}\left[\begin{array}{cc}A^{\mathrm{T}}PB & A^{\mathrm{T}}PD\end{array}\right]\left[\begin{array}{c}u_k^*\\ v_k^*\end{array}\right]+x_k^{\mathrm{T}}\bar{Q}x_k.\end{aligned} \tag{17}$$

Inserting (15) and (16) into the Hamiltonian yields the ARE in (9). This completes the proof. ■

Remark 2. *In [38], an alternative form of ARE and the optimal feedback gain K^* and L^* are described as*

$$P=A^{\mathrm{T}}\left(P^{-1}+BR_1B^{\mathrm{T}}+DR_2D^{\mathrm{T}}\right)^{-1}A+\bar{Q}, \tag{18}$$

$$K^*=-R_1^{-1}B^{\mathrm{T}}\left(P^{-1}+B^{\mathrm{T}}R_1^{-1}B^{\mathrm{T}}+D^{\mathrm{T}}R_2^{-1}D^{\mathrm{T}}\right)^{-1}A, \tag{19}$$

$$L^*=-R_2^{-1}D^{\mathrm{T}}\left(P^{-1}+B^{\mathrm{T}}R_1^{-1}B^{\mathrm{T}}+D^{\mathrm{T}}R_2^{-1}D^{\mathrm{T}}\right)^{-1}A. \tag{20}$$

It can be shown that under some manipulations, (9)-(11) are equivalent to (18)-(20). The proof of equivalence between (9)-(11) to (18)-(20) is given in APPENDIX A and B. □

As mentioned in Remark 1, the optimal control of the auxiliary system (7) with performance (8) is able to solve the robust control problem of uncertain system (1) only under some specific conditions. The condition that guarantees the feedback gain K^* in (10) asymptotically stabilizes system (1) is provided as the following theorem.

Theorem 2. *Under Assumption 2, suppose that the positive constant ε in (5) satisfies*

$$\varepsilon^{-1}I-S^{\mathrm{T}}PS\succ 0. \tag{21}$$

Then, the state feedback control $u_k=K^*x_k$ with K^* in (10) can asymptotically stabilize system (1), provided that the following is true

$$\begin{aligned}A_c^{\mathrm{T}}\left(P^{-1}-\varepsilon SS^{\mathrm{T}}\right)^{-1}A_c\prec M^{\mathrm{T}}P^{-1}M+\left(K^*\right)^{\mathrm{T}}R_1K^*\\ +\left(L^*\right)^{\mathrm{T}}R_2L^*+Q+\beta^2I,\end{aligned} \tag{22}$$

where $M=\left(P^{-1}+B^{\mathrm{T}}R_1^{-1}B+D^{\mathrm{T}}R_2^{-1}D\right)^{-1}A$ and L^* is given by (11). □

Proof: When the feedback gain K^* in (10) is applied to system (1), it can be shown that the performance function $V\left(x_k\right)$ defined in (8) is a Lyapunov function of system (1) if Assumption 2 and (22) are satisfied. First, because P is the positive definite solution of the ARE (9), then $V\left(x_k\right)=x_k^{\mathrm{T}}Px_k>0,\ x_k\neq 0$. Now it remains to show that the time difference $\Delta V\left(x_k\right)=V\left(x_{k+1}\right)-V\left(x_k\right)<0,\forall x_k\neq 0$.

Inserting the feedback gain K^* (10) into the uncertain closed-loop dynamics (6)

$$x_{k+1}=\left(A_c+S\phi\right)x_k,$$

where $A_c=A+BK^*$. The time difference of $V\left(x_k\right)$ along the state trajectory of (23) is

$$\begin{aligned}\Delta V\left(x_k\right)=x_k^{\mathrm{T}}\left(A_c^{\mathrm{T}}PA_c+\phi^{\mathrm{T}}S^{\mathrm{T}}PS\phi\right.\\ \left.+\ \phi^{\mathrm{T}}S^{\mathrm{T}}PA_c+A_c^{\mathrm{T}}PS\phi-P\right)x_k.\end{aligned} \tag{23}$$

Based on condition (21), the following is true

$$\left(\varepsilon^{-1}I-S^{\mathrm{T}}PS\right)^{-1}\succ 0.$$

Then, using the Young's inequality, one can obtain,

$$\begin{aligned}&A_c^{\mathrm{T}}PS\left(\varepsilon^{-1}I-S^{\mathrm{T}}PS\right)^{-1}S^{\mathrm{T}}PA_c\\ &+\phi^{\mathrm{T}}\left(\varepsilon^{-1}I-S^{\mathrm{T}}PS\right)\phi\\ &\geq A_c^{\mathrm{T}}PS\phi+\phi^{\mathrm{T}}S^{\mathrm{T}}PA_c.\end{aligned}$$

By rearranging items in above equation, the following is obtained

$$\begin{aligned} &A_c^{\mathrm{T}} PS\phi + \phi^{\mathrm{T}} S^{\mathrm{T}} PA_c + \phi^{\mathrm{T}} S^{\mathrm{T}} PS\phi \\ &\leq A_c^{\mathrm{T}} PS\left(\varepsilon^{-1} I - S^{\mathrm{T}} PS\right)^{-1} S^{\mathrm{T}} PA_c + \varepsilon^{-1}\phi^{\mathrm{T}}\phi. \end{aligned}$$

Inserting above equation into (23) yields

$$\begin{aligned} \Delta V(x_k) \leq x_k^{\mathrm{T}} \Big[& A_c^{\mathrm{T}} PS\left(\varepsilon^{-1} I - S^{\mathrm{T}} PS\right)^{-1} S^{\mathrm{T}} PA_c \\ & + A_c^{\mathrm{T}} PA_c + \varepsilon^{-1}\phi^{\mathrm{T}}\phi - P \Big] x_k. \end{aligned} \tag{24}$$

Based on the matrix inversion lemma [37], the following is true

$$\left(P^{-1} - \varepsilon SS^{\mathrm{T}}\right)^{-1} = P + PS\left(\varepsilon^{-1} I - S^{\mathrm{T}} PS\right)^{-1} S^{\mathrm{T}} P.$$

Then, (24) is equivalent to

$$\begin{aligned} \Delta V(x_k) \leq x_k^{\mathrm{T}} \Big[& A_c^{\mathrm{T}} \left(P^{-1} - \varepsilon SS^{\mathrm{T}}\right)^{-1} A_c \\ & + \varepsilon^{-1}\phi^{\mathrm{T}}\phi - P \Big] x_k. \end{aligned} \tag{25}$$

Replacing P in (25) with the expression in ARE (18)

$$\begin{aligned} \Delta V(x_k) \leq x_k^{\mathrm{T}} \Big[& A_c^{\mathrm{T}} \left(P^{-1} - \varepsilon SS^{\mathrm{T}}\right)^{-1} A_c + \varepsilon^{-1}\phi^{\mathrm{T}}\phi \\ - & A^{\mathrm{T}} \left(P^{-1} + B^{\mathrm{T}} R_1^{-1} B + D^{\mathrm{T}} R_2^{-1} D\right)^{-1} A - \bar{Q} \Big] x_k. \end{aligned} \tag{26}$$

Let

$$\begin{aligned} N &= P^{-1} + B^{\mathrm{T}} R_1^{-1} B + D^{\mathrm{T}} R_2^{-1} D, \\ M &= N^{-1} A, \end{aligned}$$

then,

$$\begin{aligned} & A^{\mathrm{T}} \left(P^{-1} + B^{\mathrm{T}} R_1^{-1} B + D^{\mathrm{T}} R_2^{-1} D\right)^{-1} A \\ &= A^{\mathrm{T}} N^{-1} P^{-1} N^{-1} A + A^{\mathrm{T}} N^{-1} B^{\mathrm{T}} R_1^{-1} B N^{-1} A \\ &+ A^{\mathrm{T}} N^{-1} D^{\mathrm{T}} R_2^{-1} D N^{-1} A \\ &= M^{\mathrm{T}} P^{-1} M + K^{\mathrm{T}} R_1 K + L^{\mathrm{T}} R_2 L. \end{aligned}$$

Inserting (27) into (26) yields

$$\begin{aligned} \Delta V(x_k) &= x_k^{\mathrm{T}} \left(\varepsilon^{-1}\phi^{\mathrm{T}}\phi - F\right) x_k \\ &+ x_k^{\mathrm{T}} \Big[A_c^{\mathrm{T}} \left(P^{-1} - \varepsilon SS^{\mathrm{T}}\right)^{-1} A_c - M^{\mathrm{T}} P^{-1} M \\ &- \left(K^*\right)^{\mathrm{T}} R_1 K^* - \left(L^*\right)^{\mathrm{T}} R_2 L^* - Q - \beta^2 I \Big] x_k \end{aligned} \tag{27}$$

Note that the first and second terms in (27) is negative definite if (5) in Assumption 2 and (22) holds, which also guarantees $\Delta V(x_k) < 0$. This completes the proof. ■

Remark 3. *In the proof of Theorem 2, one can observe that the parameter ε and the parameter β are used to compensate the effect of the mismatched uncertainty, Δ, on the closed-loop stability. The parameter ε should be small and the parameter β should be large to guarantee that* (21) *and* (22) *hold, respectively. Then, the robust stabilization can be guaranteed when applying the optimal solution of Problem 2 to the uncertain system* (1). □

Remark 4. *Condition (21) and (22) guarantee the asymptotic stability of system (1) when K^* serves as the state feedback gain for uncertain system (1). Note that the optimal feedback gain K^* and L^* depend on the solution of the ARE (9). In order to obtain optimal feedback gains K^* and L^*, the exact model of the auxiliary system (7) is also required for solving* (9). □

IV. On-policy Reinforcement Learning

Typically, RL approaches can be categorized into two classes: on- and off-policy [23]. On-policy RL learns the performance of the policy being carried out to the system. On the contrary, off-policy RL learns the optimal policy independently of the system's control input [25]. In this section, two variants of on-policy RL methods are developed to solve the ARE (9). Moreover, the effect of adding probing noise to the on-policy RL method is discussed.

A. Model-based On-policy RL

In this subsection, the on-policy RL-based algorithm for solving ARE (9) is discussed.

The on-policy PI starts from an admissible policy $u^0(x_k)$ and $v^0(x_k)$. In i-th iteration, the policy $u^i(x_k)$ and $v^i(x_k)$ are evaluated by solving the following on-policy Bellman equation for the value function $V^i(\cdot)$

$$\begin{aligned} & V^i(x_k) = r\left(x_k, u_k^i, v_k^i\right) + V^i(x_{k+1}) \\ &= r\left(x_k, u_k^i, v_k^i\right) + V^i\left(Ax_k + Bu^i(x_k) + Dv^i(x_k)\right) \end{aligned} \tag{28}$$

with boundary condition $V^i(0) = 0$, where $x_{k+1} = Ax_k + Bu^i(x_k) + Dv^i(x_k)$. Then, based on the value function in i-th iteration ,$V^i(\cdot)$, the iterative control law is updated as

$$\begin{aligned} & \left\{u^{i+1}(x_k), v^{i+1}(x_k)\right\} \\ &= \arg\min_{u_k, v_k} \left\{r(x_k, u_k, v_k) + V^i\left(Ax_k + Bu_k + Dv_k\right)\right\}. \end{aligned}$$

or equivalently in the form of feedback gain K^{i+1} in (29) and L^{i+1} in (30), which are shown on top of next page. The on-policy PI algorithm is summarized in Algorithm 1.

Algorithm 1 On-Policy RL without noise in the control input

1: Begin with an admissible initial control policies $u^i(\cdot)$, $v^i(\cdot)$ and set the iteration index to be $i = 0$;
2: *Policy Evaluation Step*: Evaluate policies $u^i(\cdot)$ and $v^i(\cdot)$ by solving (28) for $V^i(\cdot)$:
3: *Policy Improvement Step*: Update the iterative feedback gain K^{i+1} and L^{i+1} according to (29) and (30).
4: Let $i = i + 1$.
5: Stop if the criteria $\left\|V^i(x_k) - V^{i+1}(x_k)\right\| \leq \varepsilon$, for $\forall x_k$ is satisfied; Otherwise, go to Step 2.

The on-policy RL Algorithm 1 guarantees the convergence to the optimal value function and optimal control, i.e., $V^i(x_k) \to V^*(x_k)$, $u^i(x_k) \to u^*(x_k)$ and $v^i(x_k) \to v^*(x_k)$ as $i \to \infty$. For convergence proof, see [18] for reference.

Remark 5. *$u^i(x_k)$ and $v^i(x_k)$ can be viewed as the approximation of $u^*(x_k)$ and $v^*(x_k)$ in the i-th iteration. Note that $u^{i+1}(x_k)$ and $v^{i+1}(x_k)$ are obtained based on $V^i(x_k)$, which is the performance of $u^i(x_k)$ and $v^i(x_k)$. Therefore, in each iteration, the policy $u^i(x_k)$ and $v^i(x_k)$ have to be applied to the system in order to be improved.* □

$$K^{i+1} = -\left[R_1 + B^{\mathrm{T}}P^iB - B^{\mathrm{T}}P^iD\left(R_2 + D^{\mathrm{T}}P^iD\right)^{-1}D^{\mathrm{T}}P^iB\right]^{-1}\left[B^{\mathrm{T}}P^iA - B^{\mathrm{T}}P^iD\left(R_2 + D^{\mathrm{T}}P^iD\right)^{-1}D^{\mathrm{T}}P^iA\right] \quad (29)$$

$$L^{i+1} = -\left[R_2 + D^{\mathrm{T}}P^iD - D^{\mathrm{T}}P^iB\left(R_1 + B^{\mathrm{T}}P^iB\right)^{-1}B^{\mathrm{T}}P^iD\right]^{-1}\left[D^{\mathrm{T}}P^iA - D^{\mathrm{T}}P^iB\left(R_1 + B^{\mathrm{T}}P^iB\right)^{-1}B^{\mathrm{T}}P^iA\right] \quad (30)$$

B. Dithered On-policy RL

The trade-off between exploration and exploitation in RL is one of the critical issues with great impact on the learning performance [23], [25]. The concept of persistent excitation is closely related to the exploration in ADP [27]–[30], which guarantees the convergence of the parameter learning to the optimal case. In this subsection, we investigate the effect of probing noise on the on-policy RL algorithm.

In the policy evaluation step in Algorithm 1, the on-policy Bellman equation (28) can be equivalently written as

$$\left(x_k^{\mathrm{T}} \otimes x_k^{\mathrm{T}} - x_{k+1}^{\mathrm{T}} \otimes x_{k+1}^{\mathrm{T}}\right) vec\left(P^i\right) = r\left(x_k, u_k^i, v_k^i\right), \quad (31)$$

which is a least squares equation of $\frac{(n+1)n}{2}$ independent elements in P^i. To guarantee the existence and uniqueness of solution to (31) for online implementation, the concept of persistent excitation is required.

Definition 2. (**Persistent Excitation**) *[39] A bounded vector signal $\eta_i \in \mathbb{R}^q$, $q > 1$ is called persistently exciting (PE) if there exist $L > 0$ and $\alpha_0 > 0$ such that*

$$\sum_{i=k}^{k+L} \eta_i \eta_i^{\mathrm{T}} \geq \alpha_0 I, \forall k \geq i_0, \qquad \square$$

In order to satisfy the PE condition, a probing noise e_k is added into the control input [27]. Then, in i-th iteration, the control signal that applied to the system is

$$\bar{u}_k^i = u_k^i + e_k,$$

where e_k is a probing noise. Applying (32) to the auxiliary system (7) yields the following dithered on-policy Bellman equation

$$\begin{aligned} x_k^{\mathrm{T}}\bar{P}^i x_k &= r\left(x_k, \bar{u}_k^i, v_k^i\right) + x_{k+1}^{\mathrm{T}}\bar{P}^i x_{k+1} \\ &= x_k^{\mathrm{T}}\bar{Q}x_k + \left(\bar{u}_k^i\right)^{\mathrm{T}}R_1\bar{u}_k^i + \left(v_k^i\right)^{\mathrm{T}}R_2 v_k^i \\ &\quad + \left(Ax_k + Bu_k^i + Be_k + Dv_k^i\right)^{\mathrm{T}}\bar{P}^i \\ &\quad \left(Ax_k + Bu_k^i + Be_k + Dv_k^i\right). \end{aligned} \quad (32)$$

Based on the dithered on-policy Bellman equation (32), the on-policy PI algorithm when applying the control input with probing noise to the auxiliary system (7) is shown in Algorithm 2.

The effect of probing noise e_k on solving on-policy Bellman equation is investigated in the following lemma.

Lemma 2. *Denote the solution of the on-policy Bellman equation (28) or (31) as P^{i+1} when there is no probing noise in the control input, i.e., $e_k = 0$, and the solution of dithered on-policy Bellman equation (32) as $\bar{P}^{i+1}$ when using a probing noise in the control input, i.e., $e_k \neq 0$. Then, $P^{i+1} \neq \bar{P}^{i+1}$.*

Algorithm 2 On-Policy RL with noise in the control input

1: Set the iteration index to be $i = 0$. Begin with initial admissible control policies u_k^i, v_k^i;
2: Add probing noise e_k into the control input u_k^i to obtain $\bar{u}_k^i$. Then, apply $\bar{u}_k^i$ and v_k^i to the auxiliary system (7);
3: *Policy Evaluation Step*: Evaluate policies u_k^i and v_k^i by solving the dithered on-policy Bellman equation (32) for $\bar{P}^i$;
4: *Policy Improvement Step*: Update the iterative feedback gain K^{i+1} and L^{i+1} according to (29) and (30);
5: Let $i = i + 1$.
6: Stop if $\left\|\bar{P}^i - \bar{P}^{i+1}\right\| \leq \varepsilon$; Otherwise, go to Step 2.

Proof: Considering the auxiliary system dynamics (7), then the dithered on-policy Bellman equation (32) is equivalent to

$$\begin{aligned} &x_k^{\mathrm{T}}\bar{P}^i x_k \\ &= x_k^{\mathrm{T}}\bar{Q}x_k + \left(u_k^i\right)^{\mathrm{T}}R_1 u_k^i + \left(v_k^i\right)^{\mathrm{T}}R_2 v_k^i + x_{k+1}^{\mathrm{T}}\bar{P}^i x_{k+1} \\ &+ e_k^{\mathrm{T}}\left(R_1 + B^{\mathrm{T}}\bar{P}^i B\right)e_k + 2e_k^{\mathrm{T}}R_1 u_k^i + 2e_k^{\mathrm{T}}B^{\mathrm{T}}\bar{P}^i x_{k+1}. \end{aligned} \quad (33)$$

Note that the dithered on-policy Bellman equation (32) is the on-policy Bellman equation (28) with three extra terms related to the probing noise e_k. Then, P^{i+1}, the solution to the on-policy Bellman equation (28) does not satisfy the dithered on-policy Bellman equation (32) or (33). Therefore, P^{i+1} is not the same as $\bar{P}^{i+1}$. This completes the proof. ■

Remark 6. *From Lemma 2, it is shown that the dithered on-policy Bellman equation (32) in is inconsistent with the on-policy Bellman equation* (28)*. Therefore, Algorithm 2 will not generate the same solution as Algorithm 1. That is, Algorithm 1 is not robust to probing noise, which restricts the exploration of the on-policy RL approach.* □

Remark 7. *In both variants of on-policy RL approaches (Algorithm 1 and 2), it is shown that the policy to be evaluated has to be applied to the system. Therefore, on-policy RL is essentially an off-line algorithm. Meanwhile, in the policy evaluate step of Algorithm 1 (solving* (28) *for $V^i(\cdot)$) and Algorithm 2 (solving* (32) *for $\bar{P}^i$), the complete knowledge of system dynamics, i.e., (A, B, D), is required. Therefore, on-policy RL is a model-based method.* □

In order to obviate the off-line and model-based features of on-policy RL method, off-policy RL approach, which learns the optimal policy in a online and model-free manner, is developed in the next section.

V. Off-policy Reinforcement Learning

In this section, another type of RL methods, named off-policy RL with its variants, are developed to solve the ARE

(9), in order to obtain the robust control for Problem 1. Compared with the on-policy RL approach, it is shown that off-policy RL algorithm can solve the ARE in an online and model-free manner, while being robust to the probing noise.

A. Model-Based Off-policy RL

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

$$x_{k+1} = A^i x_k + B\left(u_k - K^i x_k\right) + D\left(v_k - L^i x_k\right), \quad i = 0, 1, 2, \cdots \tag{34}$$

where $A^i = A + BK^i + DL^i$, $u_k^i = K^i x_k, v_k^i = L^i x_k$. The policies $u(\cdot)$ and $v(\cdot)$ are the *behavior policies* that applied to the system. The policies $u_k^i = K^i x_k$ and $v_k^i = L^i x_k$ are the *iterative policies* in the learning process.

Considering the value function with respect to $u_k^i = K^i x_k$ and $v_k^i = L^i x_k$ in i-th iteration $V^i(x_k) = x_k^{\mathrm{T}} P^i x_k$, applying Taylor series expansion to the quadratic function $V^i(x_k)$ yields

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

Inserting the closed-loop system dynamics (34) gives

$$\begin{aligned} &V^i(x_k) - V^i(x_{k+1}) \\ &= x_k^{\mathrm{T}} P^i x_k - x_k^{\mathrm{T}} \left(A^i\right)^{\mathrm{T}} P^i A^i x_k \\ &- \left(u_k - K^i x_k\right)^{\mathrm{T}} B^{\mathrm{T}} P^i x_{k+1} - \left(u_k - K^i x_k\right)^{\mathrm{T}} B^{\mathrm{T}} P^i A^i x_k \\ &- \left(v_k - L^i x_k\right)^{\mathrm{T}} D^{\mathrm{T}} P^i x_{k+1} - \left(v_k - L^i x_k\right)^{\mathrm{T}} D^{\mathrm{T}} P^i A^i x_k. \end{aligned} \tag{35}$$

Based on (12), the following discrete time Lyapunov equation holds:

$$P^i = \bar{Q} + \left(K^i\right)^{\mathrm{T}} R_1 K^i + \left(L^i\right)^{\mathrm{T}} R_2 L^i + \left(A^i\right)^{\mathrm{T}} P^i A^i.$$

where $\bar{Q} = Q + F + \beta^2 I$. By taking the above equation and $V^i(x_k) = x_k^{\mathrm{T}} P^i x_k$ into (35) one can obtain the off-policy Bellman equation

$$\begin{aligned} &x_k^{\mathrm{T}} P^i x_k - x_{k+1}^{\mathrm{T}} P^i x_{k+1} \\ &= x_k^{\mathrm{T}} \bar{Q} x_k + x_k^{\mathrm{T}} \left(K^i\right)^{\mathrm{T}} R_1 K^i x_k + x_k^{\mathrm{T}} \left(L^i\right)^{\mathrm{T}} R_2 L^i x_k \\ &- \left(v_k - L^i x_k\right)^{\mathrm{T}} D^{\mathrm{T}} P^i x_{k+1} - \left(v_k - L^i x_k\right)^{\mathrm{T}} D^{\mathrm{T}} P^i A^i x_k \\ &- \left(u_k - K^i x_k\right)^{\mathrm{T}} B^{\mathrm{T}} P^i x_{k+1} - \left(u_k - K^i x_k\right)^{\mathrm{T}} B^{\mathrm{T}} P^i A^i x_k. \end{aligned} \tag{36}$$

To this end, the model-based off-policy RL for solving the off-policy Bellman equation (36) is shown in Algorithm 3.

The equivalence between the off-policy RL in Algorithm 3 and the on-policy RL in Algorithm 1 is discussed in the following lemma.

Lemma 3. *The on-policy RL in Algorithm 1 is equivalent to the off-policy RL in Algorithm 3 in the sense that the on-policy Bellman equation (28) or (31) and the off-policy Bellman equation (36) are equivalent.* □

Algorithm 3 Model-based Off-Policy RL without noise

1: Apply admissible control policies u_k and v_k to the auxiliary system (7). Let $u^i(x_k) = u_k$, $v^i(x_k) = v_k$ and set the iteration index to be $i = 0$;
2: *Policy Evaluation Step*: Evaluate policies $u^i(\cdot)$ and $v^i(\cdot)$ by solving off-policy Bellman (36) for P^i:
3: *Policy Improvement Step*: Update the iterative feedback gain K^{i+1} and L^{i+1} according to (29) and (30).
4: Let $i = i + 1$.
5: Stop if $\left\|P^i - P^{i+1}\right\| \leq \varepsilon$; Otherwise, go to Step 2.

Proof: Inserting $A^i = A + BK^i + DL^i$ into the off-policy Bellman equation (36) gives

$$\begin{aligned} &x_k^{\mathrm{T}} P^i x_k - (Ax_k + Bu_k + Dv_k)^{\mathrm{T}} P^i (Ax_k + Bu_k + Dv_k) \\ &= x_k^{\mathrm{T}} \bar{Q} x_k + x_k^{\mathrm{T}} \left(K^i\right)^{\mathrm{T}} R_1 K^i x_k + x_k^{\mathrm{T}} \left(L^i\right)^{\mathrm{T}} R_2 L^i x_k \\ &- \left(u_k - K^i x_k\right)^{\mathrm{T}} B^{\mathrm{T}} P^i (Ax_k + Bu_k + Dv_k) \\ &- \left(u_k - K^i x_k\right)^{\mathrm{T}} B^{\mathrm{T}} P^i \left(A + BK^i + DL^i\right) x_k \\ &- \left(v_k - L^i x_k\right)^{\mathrm{T}} D^{\mathrm{T}} P^i (Ax_k + Bu_k + Dv_k) \\ &- \left(v_k - L^i x_k\right)^{\mathrm{T}} D^{\mathrm{T}} P^i \left(A + BK^i + DL^i\right) x_k. \end{aligned}$$

Eliminating the common terms in the above equation yields

$$\begin{aligned} &x_k^{\mathrm{T}} P^i x_k - x_k^{\mathrm{T}} A^{\mathrm{T}} P^i A x_k \\ &= x_k^{\mathrm{T}} \bar{Q} x_k + x_k^{\mathrm{T}} \left(K^i\right)^{\mathrm{T}} R_1 K^i x_k + x_k^{\mathrm{T}} \left(L^i\right)^{\mathrm{T}} R_2 L^i x_k \\ &+ x_k^{\mathrm{T}} \left(K^i\right)^{\mathrm{T}} B^{\mathrm{T}} P^i B K^i x_k + x_k^{\mathrm{T}} \left(K^i\right)^{\mathrm{T}} B^{\mathrm{T}} P^i D L^i x_k \\ &- 2x_k^{\mathrm{T}} \left(L^i\right)^{\mathrm{T}} D^{\mathrm{T}} P^i A x_k + x_k^{\mathrm{T}} \left(L^i\right)^{\mathrm{T}} D^{\mathrm{T}} P^i B K^i x_k \\ &- 2x_k^{\mathrm{T}} \left(K^i\right)^{\mathrm{T}} B^{\mathrm{T}} P^i A x_k + x_k^{\mathrm{T}} \left(L^i\right)^{\mathrm{T}} D^{\mathrm{T}} P^i D L^i x_k. \end{aligned}$$

By rearranging terms in above equation on can obtain

$$\begin{aligned} 0 &= x_k^{\mathrm{T}} \bar{Q} x_k + x_k^{\mathrm{T}} \left(K^i\right)^{\mathrm{T}} R_1 K^i x_k + x_k^{\mathrm{T}} \left(L^i\right)^{\mathrm{T}} R_2 L^i x_k \\ &+ x_k^{\mathrm{T}} \left(A_k + BK^i + DL^i\right)^{\mathrm{T}} P^i \left(A + BK^i + DL^i\right) x_k, \\ &- x_k^{\mathrm{T}} P^i x_k \end{aligned}$$

which is equivalent to the on-policy Bellman equation (28) or (31). This completes the proof. ■

B. Dithered Model-Based Off-policy RL

Tn this subsection, the effect of the probing noise on the convergence of the off-policy RL algorithm is investigated.

Let the behavior policy with probing noise be

$$\hat{u}_k = u_k + e_k. \tag{37}$$

Considering (34), the off-policy Bellman equation for the control input $\hat{u}_k$ with the probing noise e_k can be expressed as

$$\begin{aligned} &x_k^{\mathrm{T}} \hat{P}^i x_k - [x_{k+1} + Be_k]^{\mathrm{T}} \hat{P}^i [x_{k+1} + Be_k] \\ &= x_k^{\mathrm{T}} Q x_k + x_k^{\mathrm{T}} \left(K^i\right)^{\mathrm{T}} R_1 K^i x_k + x_k^{\mathrm{T}} \left(L^i\right)^{\mathrm{T}} R_2 L^i x_k \\ &- \left(u_k + e_k - K^i x_k\right)^{\mathrm{T}} B^{\mathrm{T}} \hat{P}^i A^i x_k - \left(v_k - L^i x_k\right) D^{\mathrm{T}} \hat{P}^i A^i x_k \\ &- \left(u_k + e_k - K^i x_k\right)^{\mathrm{T}} B^{\mathrm{T}} \hat{P}^i [x_{k+1} + Be_k] \\ &- \left(v_k - L^i x_k\right) D^{\mathrm{T}} \hat{P}^i [x_{k+1} + Be_k]. \end{aligned} \tag{38}$$

Based on the dithered off-policy Bellman equation (38), the off-policy RL algorithm with probing noise in the control input is shown in Algorithm 4.

Algorithm 4 Model-based Off-Policy RL with probing noise

1: Begin with admissible policies u_k and v_k.
2: Add probing noise e_k to admissible control policy u_k to obtain $\hat{u}_k$. Apply $\hat{u}_k$ and v_k to the auxiliary system (7). Let $u^i(x_k) = u_k$, $v^i(x_k) = v_k$ and set the iteration index to be $i = 0$;
3: *Policy Evaluation Step*: Evaluate policies $\hat{u}^i(\cdot)$ and $v^i(\cdot)$ by solving dithered off-policy Bellman (38) for $\hat{P}^i$:
4: *Policy Improvement Step*: Update the iterative feedback gain K^{i+1} and L^{i+1} according to (29) and (30).
5: Let $i = i + 1$.
6: Stop if $\left\| \hat{P}^i - \hat{P}^{i+1} \right\| \leq \varepsilon$; Otherwise, go to Step 3.

The effect of probing noise e_k in (37) on solving the off-policy Bellman equation is investigated in the following lemma.

Lemma 4. *Suppose that both the off-policy Bellman equation (36) and the dithered off-policy Bellman equation (38) have unique solution. Denote the solution of the off-policy Bellman equation (36) as P^{i+1} when there is no probing noise in the control input, i.e., $e_k = 0$, and the solution of dithered off-policy Bellman equation (38) as $\hat{P}^{i+1}$ when using a probing noise in the control input, i.e., $e_k \neq 0$. Then, the off-policy Bellman equation (36) is equivalent to the dithered off-policy Bellman equation (38) in the sense that $P^{i+1} = \hat{P}^{i+1}$.* □

Proof: By expanding the terms in (38) one can obtain

$$\begin{aligned}
&x_k^{\mathrm{T}}\hat{P}^i x_k - x_{k+1}^{\mathrm{T}}\hat{P}^i x_{k+1} - 2x_{k+1}^{\mathrm{T}}\hat{P}^i B e_k - e_k^{\mathrm{T}}B^{\mathrm{T}}\hat{P}^i B e_k \\
&= x_k^{\mathrm{T}}\bar{Q}x_k + x_k^{\mathrm{T}}\left(K^i\right)^{\mathrm{T}}R_1 K^i x_k + x_k^{\mathrm{T}}\left(L^i\right)^{\mathrm{T}}R_2 L^i x_k \\
&- \left(u_k - K^i x_k\right)^{\mathrm{T}}B^{\mathrm{T}}\hat{P}^i x_{k+1} - \left(v_k - L^i x_k\right)D^{\mathrm{T}}\hat{P}^i x_{k+1} \\
&- \left(u_k - K^i x_k\right)^{\mathrm{T}}B^{\mathrm{T}}\hat{P}^i B e_k - \left(v_k - L^i x_k\right)D^{\mathrm{T}}\hat{P}^i B e_k \\
&- \left(u_k - K^i x_k\right)^{\mathrm{T}}B^{\mathrm{T}}\hat{P}^i A^i x_k - \left(v_k - L^i x_k\right)D^{\mathrm{T}}\hat{P}^i A^i x_k \\
&- x_{k+1}^{\mathrm{T}}\hat{P}^i B e_k - e_k^{\mathrm{T}}B^{\mathrm{T}}\hat{P}^i B e_k - e_k^{\mathrm{T}}B^{\mathrm{T}}\hat{P}^i A^i x_k. \qquad (39)
\end{aligned}$$

Considering the fact that

$$\begin{aligned}
x_{k+1}^{\mathrm{T}}\hat{P}^i B e_k = x_k^{\mathrm{T}}\left(A^i\right)^{\mathrm{T}}\hat{P}^i B e_k + \left(u_k - K^i x_k\right)^{\mathrm{T}}B^{\mathrm{T}}\hat{P}^i B e_k \\
+ \left(v_k - L^i x_k\right)D^{\mathrm{T}}\hat{P}^i B e_k.
\end{aligned}$$

Then, inserting the above equation into (39) yields

$$\begin{aligned}
&x_k^{\mathrm{T}}\hat{P}^i x_k - x_{k+1}^{\mathrm{T}}\hat{P}^i x_{k+1} \\
&= x_k^{\mathrm{T}}\bar{Q}x_k + x_k^{\mathrm{T}}\left(K^i\right)^{\mathrm{T}}R_1 K^i x_k + x_k^{\mathrm{T}}\left(L^i\right)^{\mathrm{T}}R_2 L^i x_k \\
&- \left(u_k - K^i x_k\right)^{\mathrm{T}}B^{\mathrm{T}}\hat{P}^i x_{k+1} - \left(u_k - K^i x_k\right)^{\mathrm{T}}B^{\mathrm{T}}\hat{P}^i A^i x_k \\
&- \left(v_k - L^i x_k\right)D^{\mathrm{T}}\hat{P}^i x_{k+1} - \left(v_k - L^i x_k\right)D^{\mathrm{T}}\hat{P}^i A^i x_k, \qquad (40)
\end{aligned}$$

which is an alternative equivalent formulation of the dithered off-policy Bellman equation (38). By comparing the dithered off-policy Bellman equation (38) or (40) with the off-policy Bellman equation (36), it can be shown that $\hat{P}^i$, the solution of the dithered off-policy Bellman equation (38), satisfies the off-policy Bellman equation (36). Therefore, $P^{i+1} = \hat{P}^{i+1}$. This completes the proof. ■

Remark 8. *According to Lemma 4, it is shown that the off-policy Bellman equation (36) is consistent with the dithered off-policy Bellman equation, i.e., Algorithm 3 is equivalent to Algorithm 4. Therefore, the probing noise added into the behavior policy will not yield a biased result for the off-policy RL algorithm. This is in contrast to the on-policy RL approaches, as discussed in Remark 6.* □

C. Model-free Off-policy RL

By using the Kronecker product, the off-policy Bellman equation (36) can be rewritten as:

$$\begin{aligned}
&\left(x_k^{\mathrm{T}}\otimes x_k^{\mathrm{T}}\right)vec\left(P^i\right) - \left(x_{k+1}^{\mathrm{T}}\otimes x_{k+1}^{\mathrm{T}}\right)vec\left(P^i\right) \\
&+ 2\left[\left(v_k - L^i x_k\right)^{\mathrm{T}}\otimes x_k^{\mathrm{T}}\right]vec\left(D^{\mathrm{T}}P^i A\right) \\
&+ \left[\left(v_k - L^i x_k\right)^{\mathrm{T}}\otimes\left(u_k + K^i x_k\right)^{\mathrm{T}}\right]vec\left(D^{\mathrm{T}}P^i B\right) \\
&+ \left[\left(v_k - L^i x_k\right)^{\mathrm{T}}\otimes\left(v_k + L^i x_k\right)^{\mathrm{T}}\right]vec\left(D^{\mathrm{T}}P^i D\right) \\
&+ 2\left[\left(u_k - K^i x_k\right)^{\mathrm{T}}\otimes x_k^{\mathrm{T}}\right]vec\left(B^{\mathrm{T}}P^i A\right) \\
&+ \left[\left(u_k - K^i x_k\right)^{\mathrm{T}}\otimes\left(u_k + K^i x_k\right)^{\mathrm{T}}\right]vec\left(B^{\mathrm{T}}P^i B\right) \\
&+ \left[\left(u_k - K^i x_k\right)^{\mathrm{T}}\otimes\left(v_k + L^i x_k\right)^{\mathrm{T}}\right]vec\left(B^{\mathrm{T}}P^i D\right) \\
&= x_k^{\mathrm{T}}\bar{Q}x_k + x_k^{\mathrm{T}}\left(K^i\right)^{\mathrm{T}}R_1 K^i x_k + x_k^{\mathrm{T}}\left(L^i\right)^{\mathrm{T}}R_2 L^i x_k. \qquad (41)
\end{aligned}$$

Let

$$\begin{aligned}
X^i = \Big[\ &\left(X_1^i\right)^{\mathrm{T}} \quad \left(X_2^i\right)^{\mathrm{T}} \quad \left(X_3^i\right)^{\mathrm{T}} \\
&\left(X_4^i\right)^{\mathrm{T}} \quad \left(X_5^i\right)^{\mathrm{T}} \quad \left(X_6^i\right)^{\mathrm{T}} \quad \left(X_7^i\right)^{\mathrm{T}}\ \Big]^{\mathrm{T}}, \qquad (42)
\end{aligned}$$

with

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

The data collected online in compact form is denoted as:

$$H_k^i = \left[\ H_{xx}^{ik} \quad H_{vx}^{ik} \quad H_{vu}^{ik} \quad H_{vv}^{ik} \quad H_{ux}^{ik} \quad H_{uu}^{ik} \quad H_{uv}^{ik}\ \right],$$

with

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

Furthermore, the utility function can be expressed in terms of the online measurement

$$\begin{aligned}
r_k^i = x_k^{\mathrm{T}}Q x_k + x_k^{\mathrm{T}}F x_k + \beta^2 x_k^{\mathrm{T}}x_k \\
+ x_k^{\mathrm{T}}\left(K^i\right)^{\mathrm{T}}R_1 K^i x_k + x_k^{\mathrm{T}}\left(L^i\right)^{\mathrm{T}}R_2 L^i x_k.
\end{aligned}$$

Finally, the Kronecker product based off-policy Bellman equation (41) can be rewritten in compact form as:

$$H_k^i X^i = r_k. \tag{43}$$

Note that in (41), there are $N = n^2+m^2+r^2+2mr+nr+mn$ unknown components. Therefore, at least N data are required to be collected in order to solve (41) or (43) by least squares methods. Assumed that $N_1 \geq N$ data are collected as

$$H_{1:N_1} X^i = \begin{bmatrix} H_1^i \\ H_2^i \\ \vdots \\ H_{N_1}^i \end{bmatrix} X^i = \begin{bmatrix} r_1 \\ r_1 \\ \vdots \\ r_{N_1} \end{bmatrix} = r_{1:N_1}. \tag{44}$$

Therefore, the least squares (LS) solution of (44)

$$\hat{X}^i = \left(H_{1:N_1}^{\mathrm{T}} H_{1:N_1}\right)^{-1} H_{1:N_1}^{\mathrm{T}} r_{1:N_1}. \tag{45}$$

Based on the least squares solution $\hat{X}^i$ in (45), the feedback gain K^i and L^i are updated as

$$\begin{aligned} K^{i+1} = &-\left[R_1 + \hat{X}_3^i + \hat{X}_6^i\left(\hat{X}_7^i + R_2\right)^{-1}\hat{X}_5^i\right]^{-1} \\ &\left[\hat{X}_2^i - \hat{X}_6^i\left(\hat{X}_7^i + R_2\right)^{-1}\hat{X}_4^i\right], \end{aligned} \tag{46}$$

$$\begin{aligned} L^{i+1} = &-\left[R_2 + \hat{X}_7^i - \hat{X}_5^i\left(R_1 + \hat{X}_3^i\right)^{-1}\hat{X}_6^i\right]^{-1} \\ &\left[\hat{X}_4^i + \hat{X}_5^i\left(R_1 + \hat{X}_3^i\right)^{-1}\hat{X}_2^i\right]. \end{aligned} \tag{47}$$

To this end, the model-free off-policy RL algorithm for solving the off-policy Bellman equation (36) or (41) is shown in Algorithm 5.

Algorithm 5 Model-free Off-Policy RL

1: *Data Collection Phase*: Apply admissible policies u_k and v_k with probing noise to the auxiliary system (7) and collect the online data $\{x_k\}$, $\{u_k\}$ and $\{v_k\}$ to form H_k^i and r_k in (43);
2: *Initialization of Learning Phase*: Set the iteration index to be $i = 0$ and initialize the iterative policies as $u^i(x_k) = u_k$, $v^i(x_k) = v_k$ to be admissible.
3: *Learning Phase 1*: Evaluate policies $u^i(\cdot)$ and $v^i(\cdot)$ by solving the LS equation (44) for $\hat{X}^i$;
4: *Learning Phase 2*: Update the iterative feedback gain K^{i+1} and L^{i+1} according to (46) and (47).
5: Let $i = i + 1$.
6: Stop if $\left\|P^i - P^{i+1}\right\| \leq \bar{\varepsilon}$, where $\bar{\varepsilon}$ is a predetermined error bound; Otherwise, go to Step 3.

Remark 9. *The solution of LS equation* (44) *in Algorithm 5 is equivalent to the off-policy Bellman equation* (36)*, i.e., Algorithm 5 is equivalent to Algorithm 3. Moreover, as shown in Lemma 4, the model-based off-policy RL approach in Algorithm 3 is robust to probing noise. Therefore, the robustness of the model-free off-policy RL method in 5 is also guaranteed.* □

Remark 10. *From the above discussions, one can observe that Algorithm 5 is equivalent to Algorithms 4 and 3, which can solve Problem 2. Based on Theorem 2, the feedback gain K^*, obtained by Algorithm 5, also solves Problem 1. Therefore, the off-policy RL algorithm together with the problem transformation provides a model-free solution to the robust stabilization problem. That is, the system matrices A, B and S are not required.* □

VI. Simulation

In this section, the on- and off-policy RL approaches are compared in terms of both the robustness against the probing noise and the dependency on the system dynamics.

Consider the discrete-time model for the rotating inverted pendulum used in [38],

$$x_{k+1} = (A + \Delta)\, x_k + B u_k, \tag{48}$$

with the nominal system drift matrix and control input dynamic matrix as

$$A = \begin{bmatrix} 1.0008 & 0.005 & 0 & 0 \\ 0.3164 & 1.008 & 0 & 0 \\ -0.0004 & 0 & 1 & 0.005 \\ -0.1666 & -0.0004 & 0 & 1 \end{bmatrix},$$
$$B = \begin{bmatrix} -0.0065 & -2.6043 & 0.0101 & 4.0210 \end{bmatrix}^{\mathrm{T}}.$$

The mismatched system uncertainty in (48) can be expressed as $\Delta = S \times \phi$, with

$$S = \begin{bmatrix} 0.0064 & -2.5648 & 0.019 & 3.9805 \end{bmatrix}^{\mathrm{T}},$$
$$\phi = p \times \sin(6k) \times \begin{bmatrix} 0.21 & 0.1 & 0.04 & 0.03 \end{bmatrix}.$$

The parameters of the uncertainty bound in (5) is selected as

$$F = \begin{bmatrix} 48.4 & 24.2 & 9.68 & 7.26 \\ 24.2 & 12.1 & 4.84 & 3.63 \\ 9.68 & 4.84 & 1.936 & 1.452 \\ 7.26 & 3.63 & 1.452 & 1.089 \end{bmatrix},$$

and $\varepsilon = 0.005$. The parameter α in the auxiliary system dynamics (7) is selected as $\alpha = 0.02$. Then, for the optimal regulation problem of the auxiliary system (7), the weight matrix is selected as $Q = \mathrm{diag}\left(\begin{bmatrix} 1 & 2 & 3 & 1 \end{bmatrix}\right)$, $R_1 = 4$, $R_2 = 3$ and $\beta = 5$. The exact solution of the ARE in (9) is

$$P^* = 10^5 \times \begin{bmatrix} 1.8279 & 0.2783 & 0.1518 & 0.1763 \\ 0.2783 & 0.0472 & 0.0263 & 0.0297 \\ 0.1518 & 0.0263 & 0.0691 & 0.0168 \\ 0.1763 & 0.0297 & 0.0168 & 0.0191 \end{bmatrix},$$

and the optimal feedback gain is

$$K^* = \begin{bmatrix} 4.0643 & 0.7396 & 0.1668 & 0.2223 \end{bmatrix},$$
$$L^* = -\begin{bmatrix} 18.6377 & 2.8882 & 1.8788 & 1.8317 \end{bmatrix}.$$

We first implement Algorithm 1 to find the solution of the ARE (9) in an off-line manner. The iterative learning process begins from the following admissible policy the auxiliary system (7)

$$K^0 = \begin{bmatrix} 4.1682 & 0.7525 & 0.1668 & 0.2296 \end{bmatrix},$$
$$L^0 = \begin{bmatrix} 15.6727 & 2.4466 & 1.5507 & 1.5519 \end{bmatrix}. \tag{49}$$

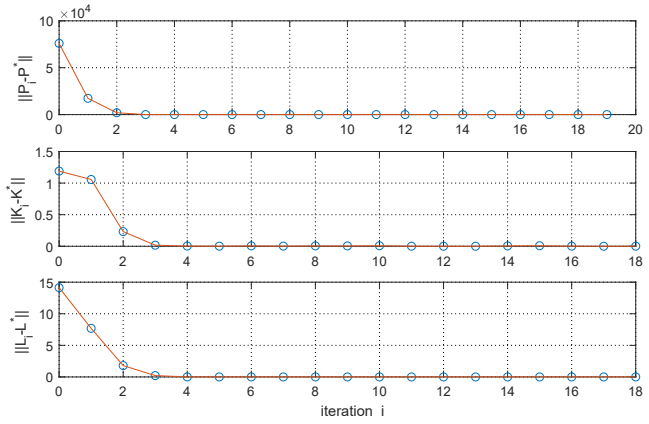

Fig. 1. Convergence of on-policy RL Algorithm 1. P^*, K^* and L^* corresponds to the optimal case, where as P_i, K_i and L_i denotes the iterative approximation in i-th iteration. The iterative learning process achieves satisfactory result after 3 iterations.

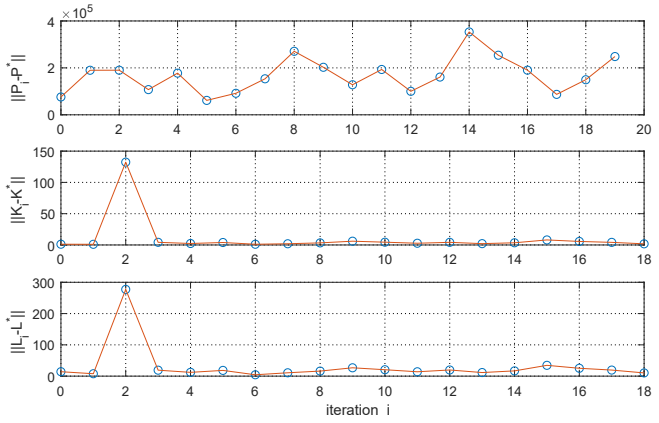

Fig. 2. Effect of probing noise on the convergence of on-policy RL algorithm. The probing noise in the iterative policy makes the learning does not converge to the optimal case.

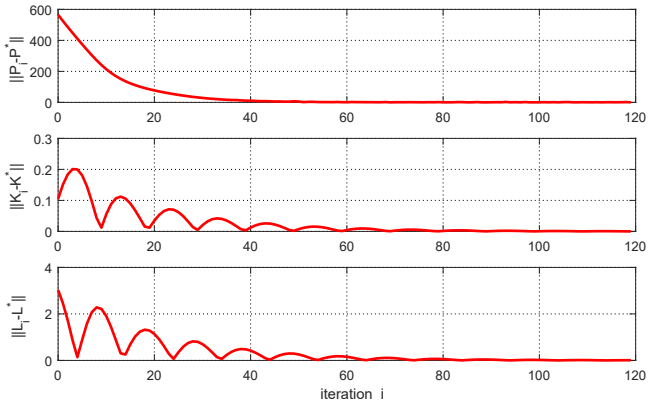

Fig. 3. Convergence of the off-policy RL algorithm with probing noise (50). P^*, K^* and L^* represents the optimal case, where as P_i, K_i and L_i denotes the learning in i-th iteration.

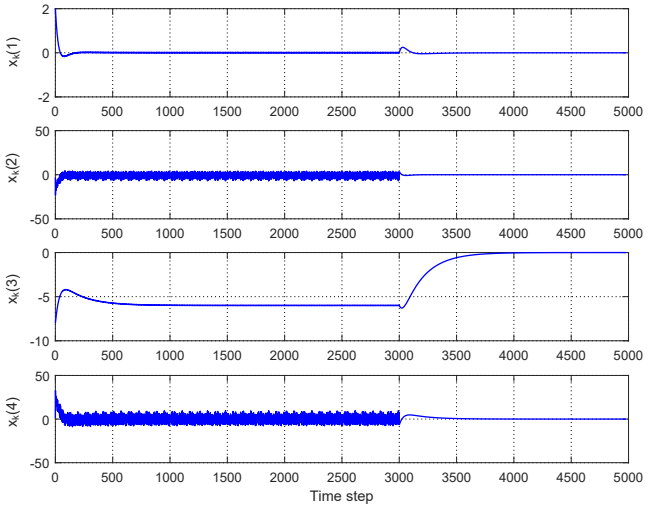

Fig. 4. The state trajectories using the off-policy RL algorithm with the probing noise (50). The system state x_k at time k is a vector which can be denoted as $x_k = [x_k(1)\ x_k(2)\ x_k(3)\ x_k(4)]^{\mathrm{T}}$, where $x_k(j)$ is the j-th element of the state x_k, for $j = 1, 2, 3, 4$.

Then, the convergence of the iterative iterative learning process of Algorithm 1 the is shown in Figure 1. Note that Algorithm 1 requires complete knowledge of the system matrices, A, B and D.

To investigate the robustness of the on-policy RL algorithm, we add the probing noise into the iterative control policy. The learning process is shown in Figure 2. As given in Lemma 2, adding the probing noise to the on-policy RL algorithm yields a bias in the learning process. One can observe that the iterative learning matrix P^i does not converge. Also, there exist nonzero residuals between the iterative feedback gains (K^i and L^i) and the optimal feedback gains (K^* and L^*). Therefore, the effect of the probing noise on the on-policy RL algorithm can not be neglected.

In the next, we implement the off-policy RL algorithm to solve the ARE (9) in an online fashion. In contrast to the on-policy RL algorithm, the behavior policy applied to the auxiliary system (7) is selected as

$$K = \begin{bmatrix} 4.1682 & 0.7525 & 0.1668 & 0.2296 \end{bmatrix},$$
$$L = \begin{bmatrix} 15.6727 & 2.4466 & 1.5507 & 1.5519 \end{bmatrix}.$$

We add two types of probing noise into the off-policy RL algorithm, i.e.,

$$e_1(k) = \cos(k) + \cos(2k) + \cos(20k),$$
$$e_2(k) = \sin(k) + \sin(0.2k). \tag{50}$$

for the first case and

$$e_1(k) = 0.1\cos(k) + \cos(2k) + \sin^2(1.7k),$$
$$e_2(k) = \sin(k) + \sin(2k) + \sin(0.538k)\cos(0.538k). \tag{51}$$

for the second case, where $e_1(k)$ is added to the iterative policy $u_k = Kx_k$ and $e_2(k)$ is added to the iterative policy $v_k = Lx_k$, respectively. The state trajectories and learning process

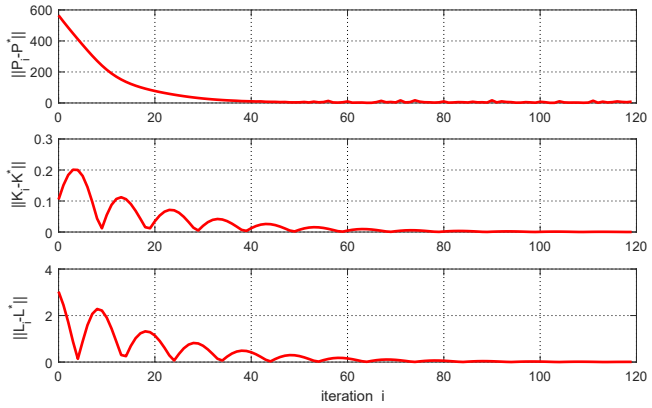

Fig. 5. Convergence of the off-policy RL algorithm with probing noise (51). P^*, K^* and L^* represents the optimal case, where as P_i, K_i and L_i denotes the learning in i-th iteration.

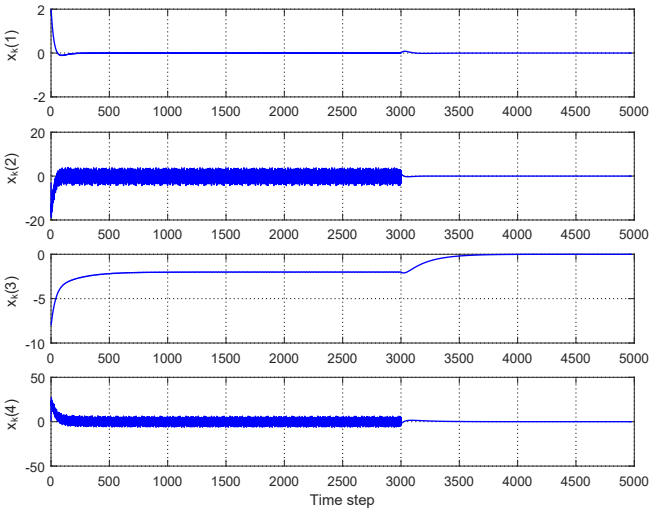

Fig. 6. The state trajectories using the off-policy RL algorithm with the probing noise (51). The system state x_k at time k is a vector which can be denoted as $x_k = [x_k(1)\ x_k(2)\ x_k(3)\ x_k(4)]^{\mathrm{T}}$, where $x_k(j)$ is the j-th element of the state x_k, for $j = 1, 2, 3, 4$.

for these two cases with different probing noises are shown in Figures 3 – 6. For the data collection phase, the probing noises is added to the behavior policy until 3000 steps. Then, the learning process begins from the same initial admissible policy for the auxiliary system (7) as given in (49). After the 3000-th step, the learning process converges to the optimal case and the approximate feedback gains K^i and L^i are implemented to the auxiliary system (7), as shown in Figures 4 and 6, respectively, which yields the asymptotically stable dynamics. The norm of learning errors between the iterative control gain K_i and K^*, L_i and L^*, between the iterative learning value function matrix P_i and optimal value function matrix P^* are shown in Figures 3 and 5, respectively. One can observe that the iterative value function matrix, P_i, the iterative control policies $u_i(x) = K_i x$ and $v_i(x) = L_i x$ converges to the solution to the ARE equation (9), P^*, the optimal control policies $u^*(x) = K^* x$ and $v^*(x) = L^* x$, respectively, as the iteration continues. In

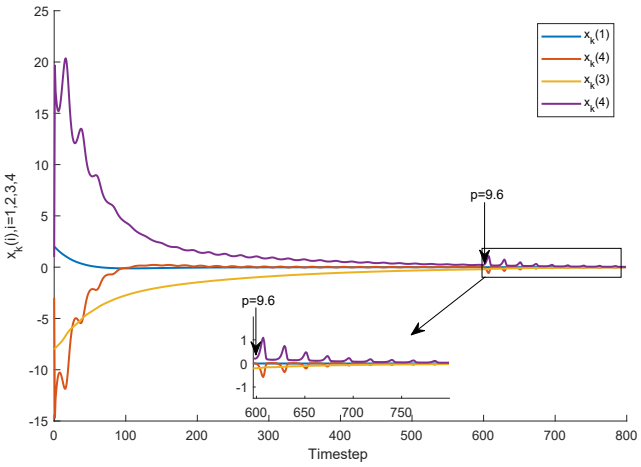

Fig. 7. Robust stabilization using the optimal controller design method. The system state x_k at time k is a vector which can be denoted as $x_k = [x_k(1)\ x_k(2)\ x_k(3)\ x_k(4)]^{\mathrm{T}}$, where $x_k(j)$ is the j-th element of the state x_k, for $j = 1, 2, 3, 4$.

addition, in both cases, both the learning errors for the gains K_i and L_i converge to the optimal gains K^* and L^*, regardless of the probing noise in the behavior policy.

When the off-policy RL algorithm converges, we use the approximate optimal feedback gain K^i to solve the robust stabilization problem. The uncertain system state trajectories are shown in Figure 7. Note that the uncertain parameter p switches from -1.6 to 9.6 at $k = 600$, which results in a small perturbation in the state trajectories. However, the robust stabilization of the closed-loop system is achieved in the sense that the state trajectories converge to the origin asymptotically, as shown in Figure 7. That is, with the presented optimal control design based method, the robust control problem of the linear dynamic system with bounded mismatched uncertainty can be solved.

VII. Conclusion

In this paper, a model-free solution is presented to solve the robust control problem of discrete-time linear dynamical systems. The robust control problem is first translated into an optimal control problem with sufficient condition which guarantees the equivalence of problem translation. Then, off-policy reinforcement learning (RL) is used to solve the translated optimal control problem using only measured data instead of the system dynamics. Moreover, compared with the on-policy RL method, it is shown theoretically that the off-policy RL method has two main advantages. First, off-policy is robust to the probing noise, i.e., there is no bias as a result of adding a probing noise to the control input to satisfy the condition of persistence of excitation. In addition, off-policy RL is a model-free method, which is in contrast to the model-based on-policy RL method. Finally, a simulation example is given to verify the effectiveness of the presented off-policy RL algorithm.

Acknowledgment

The authors would like to thank Dr. Hamidreza Modares at Michigan State University for his insightful discussions about

some of the issues addressed in this paper and the anonymous reviewers for their valuable comments and suggestions to improve the quality of this paper.

APPENDIX A
PROOF OF THE EQUIVALENCE BETWEEN (9) AND (18)

First, $\left(P^{-1}+B^{\mathrm{T}}R_1^{-1}B^{\mathrm{T}}+D^{\mathrm{T}}R_2^{-1}D^{\mathrm{T}}\right)^{-1}$ can be rewritten as

$$\begin{aligned}&\left(P^{-1}+B^{\mathrm{T}}R_1^{-1}B^{\mathrm{T}}+D^{\mathrm{T}}R_2^{-1}D^{\mathrm{T}}\right)^{-1}\\&=\left(P^{-1}-\begin{bmatrix}B^{\mathrm{T}}\\D^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}-R_1&0\\0&-R_2\end{bmatrix}^{-1}\begin{bmatrix}B^{\mathrm{T}}\\D^{\mathrm{T}}\end{bmatrix}\right)^{-1}.\end{aligned}$$

Based on the matrix inversion lemma [37], the above equation is equivalent to

$$\begin{aligned}&\left(P^{-1}-\begin{bmatrix}B^{\mathrm{T}}\\D^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}-R_1&0\\0&-R_2\end{bmatrix}^{-1}\begin{bmatrix}B^{\mathrm{T}}\\D^{\mathrm{T}}\end{bmatrix}\right)^{-1}\\&=P-P\begin{bmatrix}B^{\mathrm{T}}\\D^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}\left(\begin{bmatrix}-R_1&0\\0&-R_2\end{bmatrix}\right.\\&\quad\left.-\begin{bmatrix}B^{\mathrm{T}}\\D^{\mathrm{T}}\end{bmatrix}P\begin{bmatrix}B^{\mathrm{T}}\\D^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}\right)^{-1}\begin{bmatrix}B^{\mathrm{T}}\\D^{\mathrm{T}}\end{bmatrix}P\\&=P-\begin{bmatrix}B^{\mathrm{T}}P\\D^{\mathrm{T}}P\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}R_1+B^{\mathrm{T}}PB&B^{\mathrm{T}}PD\\D^{\mathrm{T}}PB&R_2+D^{\mathrm{T}}PD\end{bmatrix}^{-1}\\&\quad\begin{bmatrix}B^{\mathrm{T}}P\\D^{\mathrm{T}}P\end{bmatrix}\end{aligned}\tag{52}$$

By multiplying A^{T} and A on both sides of (52), then adding $\bar{Q}$ yields the equivalence between (9) and (18).

APPENDIX B
PROOF OF THE EQUIVALENCE BETWEEN (10), (11) AND (19), (20)

Based on (15), the following holds

$$\begin{aligned}&\begin{bmatrix}K^*\\L^*\end{bmatrix}=-\mathcal{M}^{-1}\begin{bmatrix}\mathcal{E}\\\mathcal{G}\end{bmatrix}\\&=-\begin{bmatrix}R_1^{-1}&0\\0&R_2^{-1}\end{bmatrix}\begin{bmatrix}R_1&0\\0&R_2\end{bmatrix}\mathcal{M}^{-1}\begin{bmatrix}\mathcal{E}\\\mathcal{G}\end{bmatrix}.\end{aligned}\tag{53}$$

Note that

$$\begin{bmatrix}R_1&0\\0&R_2\end{bmatrix}=\mathcal{M}-\begin{bmatrix}B^{\mathrm{T}}\\D^{\mathrm{T}}\end{bmatrix}P\begin{bmatrix}B&D\end{bmatrix}.\tag{54}$$

Inserting (54) into (53) yields

$$\begin{aligned}\begin{bmatrix}K^*\\L^*\end{bmatrix}&=-\begin{bmatrix}R_1^{-1}&0\\0&R_2^{-1}\end{bmatrix}\left(\mathcal{M}-\begin{bmatrix}B^{\mathrm{T}}\\D^{\mathrm{T}}\end{bmatrix}P\begin{bmatrix}B&D\end{bmatrix}\right)\\&\quad\mathcal{M}^{-1}\begin{bmatrix}\mathcal{E}\\\mathcal{G}\end{bmatrix}\\&=-\begin{bmatrix}R_1^{-1}&0\\0&R_2^{-1}\end{bmatrix}\left(\begin{bmatrix}\mathcal{E}\\\mathcal{G}\end{bmatrix}-\begin{bmatrix}B^{\mathrm{T}}\\D^{\mathrm{T}}\end{bmatrix}P\right.\\&\quad\left.\begin{bmatrix}B&D\end{bmatrix}\mathcal{M}^{-1}\begin{bmatrix}\mathcal{E}\\\mathcal{G}\end{bmatrix}\right)\\&=-\begin{bmatrix}R_1^{-1}B^{\mathrm{T}}PA\\R_2^{-1}D^{\mathrm{T}}PA\end{bmatrix}+\begin{bmatrix}R_1^{-1}B^{\mathrm{T}}P\\R_2^{-1}D^{\mathrm{T}}P\end{bmatrix}\begin{bmatrix}B&D\end{bmatrix}\\&\quad\begin{bmatrix}R_1+B^{\mathrm{T}}PB&B^{\mathrm{T}}PD\\D^{\mathrm{T}}PB&R_2+D^{\mathrm{T}}PD\end{bmatrix}^{-1}\begin{bmatrix}B^{\mathrm{T}}PA\\D^{\mathrm{T}}PA\end{bmatrix}.\end{aligned}\tag{55}$$

The first row of (55) gives

$$\begin{aligned}K^*&=-R_1^{-1}B^{\mathrm{T}}PA+R_1^{-1}B^{\mathrm{T}}P\begin{bmatrix}B&D\end{bmatrix}\\&\quad\begin{bmatrix}R_1+B^{\mathrm{T}}PB&B^{\mathrm{T}}PD\\D^{\mathrm{T}}PB&R_2+D^{\mathrm{T}}PD\end{bmatrix}^{-1}\begin{bmatrix}B^{\mathrm{T}}\\D^{\mathrm{T}}\end{bmatrix}PA\\&=-R_1^{-1}B^{\mathrm{T}}\left\{P-P\begin{bmatrix}B&D\end{bmatrix}\left(\begin{bmatrix}R_1&0\\0&R_2\end{bmatrix}\right.\right.\\&\quad\left.\left.+\begin{bmatrix}B^{\mathrm{T}}\\D^{\mathrm{T}}\end{bmatrix}P\begin{bmatrix}B&D\end{bmatrix}\right)^{-1}\begin{bmatrix}B^{\mathrm{T}}\\D^{\mathrm{T}}\end{bmatrix}P\right\}A\\&=-R_1^{-1}B^{\mathrm{T}}\left(P^{-1}+B^{\mathrm{T}}R_1^{-1}B+D^{\mathrm{T}}R_2^{-1}D\right)^{-1}A.\end{aligned}$$

Therefore, (19) is equivalent to (10). The equivalence between (20) and (11) can be obtained in a similar way.

REFERENCES

[1] K. Zhou and J. C. Doyle, *Essentials of robust control*. Upper Saddle River, NJ, USA: Prentice-Hall, 1998.
[2] G. E. Dullerud and F. Paganini, *A course in robust control theory: a convex approach*. New York, NY, USA: Springer, 2013, vol. 36.
[3] K.-Z. Liu and Y. Yao, *Robust Control: Theory and Applications*. Solaris South Tower, Singapore: John Wiley & Sons, 2016.
[4] X. Xie, D. Yue, H. Zhang, and C. Peng, "Control synthesis of discrete-time ts fuzzy systems: Reducing the conservatism whilst alleviating the computational burden," *IEEE Transactions on Cybernetics*, vol. 47, no. 9, pp. 2480–2491, Sept 2017.
[5] X. Xie, D. Yue, and C. Peng, "Relaxed real-time scheduling stabilization of discrete-time takagi-sugeno fuzzy systems via a alterable-weights-based ranking switching mechanism," *IEEE Transactions on Fuzzy Systems*, pp. 1–1, 2018.
[6] J. Doyle and G. Stein, "Multivariable feedback design: Concepts for a classical/modern synthesis," *IEEE Transactions on Automatic Control*, vol. 26, no. 1, pp. 4–16, Feb 1981.
[7] G. Zames, "Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses," *IEEE Transactions on Automatic Control*, vol. 26, no. 2, pp. 301–320, April 1981.
[8] S. Boyd and C. Barratt, *Linear controller design: limits of performance*. Upper Saddle River, NJ, USA: Prentice-Hall, 1991.
[9] J. C. Doyle, K. Glover, P. P. Khargonekar, and B. A. Francis, "State-space solutions to standard H_2 and H_∞ control problems," *IEEE Transactions on Automatic Control*, vol. 34, no. 8, pp. 831–847, Aug 1989.
[10] F. Lin, *Robust control design: an optimal control approach*. John Wiley & Sons, 2007, vol. 18.
[11] W. B. Powell, *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Hoboken, NJ, USA: Wiley, 2007.

[12] P. J. Werbos, "Approximate dynamic programming for real-time control and neural modeling," in *Handbook of intelligent control*, D. A. White and D. A. Sofge, Eds. New York, NY, USA: Van Nostrand Reinhold Company, 1992, pp. 493–526.
[13] ——, *The roots of backpropagation: from ordered derivatives to neural networks and political forecasting*. John Wiley & Sons, 1994.
[14] D. V. Prokhorov and D. C. Wunsch, "Adaptive critic designs," *IEEE Transactions on Neural Networks*, vol. 8, no. 5, pp. 997–1007, Sep 1997.
[15] H. He, Z. Ni, and J. Fu, "A three-network architecture for on-line learning and optimization based on adaptive dynamic programming," *Neurocomputing*, vol. 78, no. 1, pp. 3 – 13, 2012.
[16] Z. Ni, H. He, and J. Wen, "Adaptive learning in tracking control based on the dual critic network design," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 6, pp. 913–928, June 2013.
[17] Y. Yang, D. Wunsch, and Y. Yin, "Hamiltonian-driven adaptive dynamic programming for continuous nonlinear dynamical systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 8, pp. 1929–1940, Aug 2017.
[18] D. Liu and Q. Wei, "Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 3, pp. 621–634, March 2014.
[19] K. G. Vamvoudakis and F. L. Lewis, "Online actorcritic algorithm to solve the continuous-time infinite horizon optimal control problem," *Automatica*, vol. 46, no. 5, pp. 878 – 888, 2010.
[20] S. Bhasin, R. Kamalapurkar, M. Johnson, K. Vamvoudakis, F. Lewis, and W. Dixon, "A novel actorcriticidentifier architecture for approximate optimal control of uncertain nonlinear systems," *Automatica*, vol. 49, no. 1, pp. 82 – 92, 2013.
[21] H. Modares, F. L. Lewis, and M. B. Naghibi-Sistani, "Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 10, pp. 1513–1525, Oct 2013.
[22] R. Kamalapurkar, L. Andrews, P. Walters, and W. E. Dixon, "Model-based reinforcement learning for infinite-horizon approximate optimal tracking," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 3, pp. 753–758, March 2017.
[23] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.
[24] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis, "Optimal and autonomous control using reinforcement learning: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2042–2062, June 2018.
[25] D. L. Poole and A. K. Mackworth, *Artificial Intelligence: foundations of computational agents*. Cambridge University Press, 2010.
[26] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3, pp. 279–292, May 1992.
[27] Y. Jiang and Z.-P. Jiang, "Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics," *Automatica*, vol. 48, no. 10, pp. 2699–2704, 2012.
[28] H. Modares, F. L. Lewis, and Z. P. Jiang, "H_∞ 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, Oct 2015.
[29] B. Luo, H. N. Wu, and T. Huang, "Off-policy reinforcement learning for H_∞ control design," *IEEE Transactions on Cybernetics*, vol. 45, no. 1, pp. 65–76, Jan 2015.
[30] 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.
[31] Y. Yang, H. Modares, D. C. Wunsch, and Y. Yin, "Leaderfollower output synchronization of linear heterogeneous systems with active leader using reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2139–2153, June 2018.
[32] H. Modares, S. P. Nageshrao, G. A. D. Lopes, R. Babuka, and F. L. Lewis, "Optimal model-free output synchronization of heterogeneous systems using off-policy reinforcement learning," *Automatica*, vol. 71, pp. 334 – 341, 2016.
[33] B. Kiumarsi and F. L. Lewis, "Output synchronization of heterogeneous discrete-time systems: A model-free optimal approach," *Automatica*, vol. 84, pp. 86 – 94, 2017.
[34] Y. Yang, H. Modares, D. C. Wunsch, and Y. Yin, "Optimal containment control of unknown heterogeneous systems with active leaders," *IEEE Transactions on Control Systems Technology*, vol. PP, no. 99, pp. 1–9, 2018. [Online]. Available: http://ieeexplore.ieee.org/document/8277155/
[35] K. G. Vamvoudakis, H. Modares, B. Kiumarsi, and F. L. Lewis, "Game theory-based control system algorithms with real-time reinforcement learning: How to solve multiplayer games online," *IEEE Control Systems*, vol. 37, no. 1, pp. 33–52, Feb 2017.
[36] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal control*, 3rd ed. Hoboken, NJ, USA: John Wiley & Sons, 2012.
[37] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.
[38] N. S. Tripathy, I. N. 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, Sept 2017.
[39] G. Tao, *Adaptive Control Design and Analysis*. Hoboken, New Jersey, USA: John Wiley & Sons, 2003.

Yongliang Yang (M'18) received the B.S. degree from Hebei University, Baoding, China, in 2011 and the Ph.D. degree from the University of Science and Technology Beijing (USTB), Beijing, China, in 2017. He was a visiting scholar with the Missouri University of Science and Technology, Rolla, USA from 2015 to 2017, sponsored by China Scholarship Council. He was the recipient Best Ph.D. Dissertation, Chancellor's Scholarship, Scholarship for Outstanding Ph.D. Students in USTB, and Excellent Graduates Awards in Beijing.

Dr. Yang is currently an Assistant Professor at USTB. He also serves as the reviewer of several international journals and conferences, including IEEE Transactions on Automatic Control, Automatica, IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Cybernetics, IEEE Transactions on Systems, Man, and Cybernetics: Systems, Neurocomputing and IEEE/CAA Journal of Automatica Sinica. His research interests include adaptive optimal control, distributed optimization and control, and cyber-physical systems.

Zhishan Guo (M'15) received the B.E. degree in computer science and technology from Tsinghua University, Beijing, China, in 2009, and the M. Phil. degree in mechanical and automation engineering from the Chinese University of Hong Kong, Hong Kong, in 2011.

Dr. Guo is currently an assistant professor in the Department of Electric and Computer Engineering, University of Central Florida, Orlando, Florida, USA. His research and teaching interests include real-time scheduling, cyber-physical systems, and neural networks and their applications.

Haoyi Xiong (M'15) received the PhD degree with highest honors in computer science from Telecom SudParis and the University of Paris VI, France, in 2015. He was an assistant professor in the Department of Computer Science, Missouri University of Science & Technology, in Rolla, Missouri (previously known as University of Missouri at Rolla) from 2016 to 2018. From July 2015 to August 2016, he was a postdoctoral research associate in the Department of Systems and Information Engineering, University of Virginia, Charlottesville, Virginia. He is currently with the Big Data Laboratory, Baidu Research, Beijing, China.

His research interests include ubiquitous computing, cyber-human systems, and large-scale optimization & decision making. He received the Best Paper Award from IEEE UIC 2012 and the Outstanding PhD thesis Runner-up Award from CNRS SAMOVAR 2015. He is a member of the IEEE.

Da-Wei Ding received the B.E. degree from the Ocean University of China, Qingdao, China, in 2003, and the Ph.D. degree in control theory and engineering from Northeastern University, Shenyang, China, in 2010. He is currently a Professor with the School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing, China. His research interests include robust control and filtering, fuzzy control, and fault detection.

Yixin Yin (M'07) received the B.S., M.S., and Ph.D. degrees from the University of Science and Technology Beijing (USTB), Beijing, China, in 1982, 1984, and 2002, respectively. He is the recipient of several national awards, including the Outstanding Young Educator in 1993, the Award of Science and Technology Progress in Education in 1994, the Award of National Science and Technology Progress in 1995, the Special Allowance of the State Council in 1994, the Best Paper of Japanese Acoustical Society in 1999, and the Award of Metallurgical Science and Technology in 2014. He was a visiting scholar with several universities in Japan, including University of Tokyo, Kyushu Institute of Technology, Kanagawa University, Chiba University and Muroran Institute of Technology.

Prof. Yin served as the dean of School of the Information Engineering, USTB from 2000 to 2011, and the dean of the School of Automation and Electrical Engineering, USTB from 2011 to 2017. He is a Fellow of Chinese Society for Artificial Intelligence, a Member of Chinese Society for Metals and Chinese Association of Automation. He is currently a professor with the School of Automation and Electrical Engineering, USTB. His major research interests include modeling and control of complex industrial processes, computer aided design of control system, intelligent control and artificial life.

Donald Wunsch (F'05) is the Mary K. Finley Missouri Distinguished Professor at Missouri University of Science and Technology (Missouri S&T), Rolla, Missouri. He is the Director of the Applied Computational Intelligence Laboratory, a multidisciplinary research group. Earlier employers were: Texas Tech University (Lubbock, TX,) Boeing (Seattle, WA,) Rockwell International (Albuquerque, NM,) and International Laser Systems, (Albuquerque, NM). His education includes: Ph.D., Electrical Engineering - University of Washington (Seattle), Executive MBA - Washington University in St. Louis, M.S., Applied Mathematics - University of Washington (Seattle), B.S., Applied Mathematics - University of New Mexico (Albuquerque, NM), and Jesuit Core Honors Program, Seattle University. Key research contributions are: Clustering / Unsupervised Learning; Biclustering; Adaptive Resonance and Adaptive Dynamic Programming architectures, hardware and applications; Neurofuzzy regression; Autonomous Agents; Games; and Bioinformatics.

Prof. Wunsch is an IEEE Fellow and previous International Neural Networks Society (INNS) President, INNS Fellow, NSF CAREER Awardee, 2015 INNS Gabor Award recipient, and 2019 Ada Lovelace Service Award Recipient. He served as IJCNN General Chair, and on several Boards, including the St. Patricks School Board, IEEE Neural Networks Council, INNS, and the University of Missouri Bioinformatics Consortium, Chaired the Missouri S&T Information Technology and Computing Committee as well as the Student Design and Experiential Learning Center Board. He has produced 21 Ph.D. recipients in Computer Engineering, Electrical Engineering, Systems Engineering and Computer Science.