J Syst Sci Complex (2023) 36: 591–612

Optimal Control of Unknown Discrete-Time Linear Systems with Additive Noise*

YANG Xue · LIU Shujun

DOI: 10.1007/s11424-023-1352-4
Received: 18 March 2021 / Revised: 16 September 2021

Abstract The optimal control problem with a long run average cost is investigated for unknown linear discrete-time systems with additive noise. The authors propose a value iteration-based stochastic adaptive dynamic programming (VI-based SADP) algorithm, based on which the optimal controller is obtained. Different from the existing relevant work, the algorithm does not need to estimate the expectation (conditional expectation) and variance (conditional variance) of states or other relevant variables, and the convergence of the algorithm can be proved rigorously. A simulation example is given to verify the effectiveness of the proposed approach.

Keywords Discrete-time linear systems, optimal control, stochastic adaptive dynamic programming.

1 Introduction

Optimal control plays an essential role in modern control theory[1]. The majority of previous optimal control methods require perfect knowledge of system dynamics. However, in the real world, formulating an accurate mathematical dynamic model is hard. To tackle this difficulty, some methods are proposed to solve optimal control and related problems of unknown systems (e.g., [2–6] and the references therein).

For deterministic discrete-time linear systems, Kiumarsi, et al.[7] solved an H_∞ control problem by applying an off-policy reinforcement learning (RL) algorithm; Lewis and Vamvoudakis[8] developed both policy iteration and value iteration-based adaptive dynamic programming (PI and VI-based ADP) algorithms to solve the linear quadratic (LQ) regulation problem; Rizvi and Lin[9] also proposed both PI and VI-based Q-learning approaches to solve the LQ regulation problem without resorting to a discounting factor; Kiumarsi, et al.[10] gave a PI-based Q-learning RL algorithm for the LQ optimal tracking control problem; Jiang, et al.[11] developed off-line PI, online PI, and Q-learning PI RL algorithms for the optimal tracking control

YANG Xue · LIU Shujun (Corresponding author)
School of Mathematics, Sichuan University, Chengdu 610064, *China.* Email: sjliu@scu.edu.cn.
*This research was supported by the National Natural Science Foundation of China under Grant No. 61673284 and the Science Development Project of Sichuan University under Grant No. 2020SCUNL201.
◇ *This paper was recommended for publication by Editor WANG Le.*

problem of networked control systems with dropout. For deterministic discrete-time nonlinear systems, He and Jagannathan[12] designed a neural network-based (NN-based) RL algorithm to solve the optimal tracking control problem with input constraints; Wei and Liu[13] developed a PI-based deterministic Q-learning algorithm to solve the optimal control problems via NN; Wang, et al.[14] proposed a PI-based ADP approach via NN for robust control problem.

Actual systems are usually affected by noise. Generally, stochastic models are more in line with practical problems than deterministic ones. Recently, the optimal control and related problems of unknown stochastic discrete-time ones have achieved some attentions both in theoretical methods ([15–25]) and in applications ([26–28]). For discrete-time linear systems with multiplicative noise, Liu, et al.[15] solved LQ optimal control problem of unknown mean-field discrete-time systems based on VI ADP via NN; Liu, et al.[16] solved the LQ Stackelberg game problem by VI-based ADP via NN; Gravell, et al.[17] gave a PI-based ADP algorithm for stochastic LQ zero-sum game problem. For discrete-time linear systems with additive noise, Wang and Yang[18] obtained nearly optimal control laws for the optimal control problem by VI-based ADP method; Han, et al.[19] developed a PI-based ADP technique to solve the fault-tolerant optimal tracking control problem; Wong and Lee[20] solved an optimal control problem by establishing a PI-based Q-learning RL algorithm; Yaghmaie and Gustafsson[21] developed an off-policy PI-based RL method for the LQ optimal control problem; Abbasi-Yadkori, et al.[22] proposed a new model-free algorithm via reduction to expert prediction for LQ optimal control problem. For discrete-time stochastic nonlinear systems, Xu, et al.[23] developed a learning-based predictive control algorithm via NN for the optimal control; Liang, et al.[24, 25] respectively proposed a local PI and an improved VI ADP algorithm via NN to slove the optimal control problem under the assumption that the past and the current system noises are independent and the state space and the admissible control set are countable.

However, most of these work make use of the expectation (conditional expectation) and variance (conditional variance) of states or other relevant variables in the algorithm design (e.g., [15–25]), which may result in some limitations or disadvantages: 1) Many trajectory information or data are needed to estimate the expectation and variance; 2) the control policy designed is not optimal owing to the existence of the estimate error; 3) the convergence of the algorithm or the optimality of the controller is hard to analyze or prove rigorously.

In this paper, we investigate the long run average optimal control problem for unknown linear discrete-time systems with additive noise. The performance index is a long time-averaged cost which has a broad practical background[29, 30]. To tackle this problem, we propose a value iteration-based stochastic adaptive dynamic programming (VI-based SADP) algorithm. We use the stochastic recursive least squares method to obtain the estimated value of the relevant parameters. Compared with existing methods[15–25], our proposed algorithm can obtain the optimal controller by directly using the value of the system states without estimating the expectation (conditional expectation) and variance (conditional variance) of states or other variables. The system state space and the admissible control set can be uncountable in our method. Different from [19–22], the coefficient matrix of the additive noise term can be any unknown matrix instead of a known or unit matrix.

It should be pointed out that there are some similar works[31–33] about stochastic optimal control based on system parameter estimation. Compared with [31], each value iteration in this paper does not need to solve the algebraic Riccati equation. Different from the autoregressive-moving average with exogenous input model in [32, 33], the state space model is considered in this paper, and moreover, the autoregressive-moving average with exogenous input model can be transformed into a state space model, where the type of coefficient matrices is a special case in our model. These similar works use a kind of indirect adaptive optimal control methods, in which controls are recomputed from an estimated system model at each step and this computation is inherently complex[34].

The contributions of our work includes: 1) We propose a value iteration-based stochastic adaptive dynamic programming algorithm, based on which the optimal controller is obtained; 2) the convergence of the value iteration-based stochastic adaptive dynamic programming algorithm and the optimality of the controller can be proved rigorously under some reasonable conditions; 3) we directly use the value of system states without estimating the expectation (conditional expectation) and variance (conditional variance) of states or other variables in the algorithm design.

The remainder of the paper is organized as follows. In Section 2, we present the formulation of the optimal control problem and review the optimal control method for known discrete-time systems. In Section 3, we give the VI-based SADP design for unknown linear discrete-time systems with additive noise. In Section 4, we provide the VI-based SADP algorithm, the optimal controller, the algorithm's convergence result, and the optimal result of the control policy. In Section 5, we offer a simulation example to show the effectiveness of the algorithm and the controller. In Section 6, we give some concluding remarks.

Notations $\|\cdot\|$ represents the Euclidean norm for vectors, or the induced matrix norm for matrices. $\boldsymbol{S}_r$ is the normed space with the induced matrix norm of all r-by-r real symmetric matrices. $\mathbb{N}$ denotes the set of natural numbers. $E(\cdot)$ denotes mathematical expectation. For any $M \in \boldsymbol{S}_r$, denote $\lambda_{\min}(M)$ as the minimum eigenvalue of M. F^+ represents the pseudo-inverse of the matrix F. For a symmetric matrix $P = (p_{ij}) \in \mathbb{R}^{m\times m}$, define vector-valued function $\mathrm{vecs}(P) = [p_{11}\ \ 2p_{12}\ \ \cdots\ 2p_{1m}\ \ p_{22}\ \ 2p_{23}\ \ \cdots\ \ 2p_{m-1,m}\ \ p_{mm}]^{\mathrm{T}} \in \mathbb{R}^{\frac{1}{2}m(m+1)}$. For an arbitrary column vector $v = [v_1\ \ v_2\ \ \cdots\ \ v_n]^{\mathrm{T}} \in \mathbb{R}^n$, vector-valued function $\mathrm{vecv}(v) = [v_1^2\ \ v_1v_2\ \ \cdots\ \ v_1v_n\ \ v_2^2\ \ v_2v_3\ \ \cdots\ \ v_{n-1}v_n\ \ v_n^2]^{\mathrm{T}} \in \mathbb{R}^{\frac{1}{2}n(n+1)}$. $\otimes$ indicates the Kronecker product operator and $\mathrm{vec}(\widehat{A}) = [\widehat{a}_1^{\mathrm{T}}\ \ \widehat{a}_2^{\mathrm{T}}\ \ \cdots\ \ \widehat{a}_m^{\mathrm{T}}]^{\mathrm{T}}$, where $\widehat{a}_i \in \mathbb{R}^n$ are the columns of $\widehat{A} \in \mathbb{R}^{n\times m}$.

2 Problem Formulation

Consider the following discrete time stochastic system

$$x_{k+1} = Ax_k + Bu_k + C\omega_{k+1}, \tag{1}$$

where $x_k \in \mathbb{R}^r$ is the system state, $u_k \in \mathbb{R}^{r_1}$ is the control input, and $\{\omega_k \in \mathbb{R}^d, k = 0, 1, \cdots\}$ is the noise defined in a probability space $(\Omega, \mathcal{F}, P)$ with respect to a nondecreasing family of σ-algebras $\{\mathcal{F}_k\}$. The coefficients A, B, and C are assumed to be unknown deterministic

matrices with appropriate dimensions.

Consider the long run average cost function

$$J(u)=\limsup_{k\to\infty}\frac{1}{k}\sum_{i=0}^{k-1}(x_i^{\mathrm{T}}Qx_i+u_i^{\mathrm{T}}Ru_i), \tag{2}$$

where $Q=Q^{\mathrm{T}}>0$ and $R=R^{\mathrm{T}}>0$.

The admissible control set is[36]

$$\mathcal{U}_{ad}=\left\{u|\ \|x_k\|^2=o(k),\ \sum_{i=0}^{k-1}(\|u_i\|^2+\|x_i\|^2)=O(k)\ a.s.,\ u_i\in\mathcal{F}_i,\ i\geq 0\right\}. \tag{3}$$

We consider the following assumptions.

(A1) $x_0\in\mathcal{F}_0$.

(A2) (A,B) is controllable and (A,D) is observable, where D is any matrix satisfying $D^TD=Q$.

(A3) $\{\omega_k,\mathcal{F}_k\}_{k=0}^{\infty}$ is an almost surely bounded martingale difference sequence, $\omega_{k,1}$, $\omega_{k,2}$, $\cdots$, $\omega_{k,d}$ are mutually independent for give k, and

$$\begin{cases}E(\omega_{k,s}^3|\mathcal{F}_{k-1})=0\quad\text{a.s.},\\ E\{\omega_{k,s}^2|\mathcal{F}_{k-1}\}=\xi_s^2>0\quad\text{a.s.},\\ \lim\limits_{k\to\infty}\dfrac{1}{k}\sum\limits_{i=1}^{k}\omega_i\omega_i^{\mathrm{T}}=V>0\quad\text{a.s.},\\ \lim\limits_{k\to\infty}\dfrac{1}{k}\sum\limits_{i=0}^{k-1}(\omega_{i,s}^2-\xi_s^2)^2\neq 0\quad\text{a.s.},\end{cases} \tag{4}$$

for $s=1,2,\cdots,d$, where $\omega_k=[\omega_{k,1},\omega_{k,2},\cdots,\omega_{k,d}]^{\mathrm{T}}$ and $V\in\mathbb{R}^{d\times d}$ is a deterministic matrix.

(A4) The sign of a certain non-zero element in matrices A and B is known, and $\det(A)\neq 0$.

(A5) The coefficient matrix C of additive noise is of row full rank.

Remark 2.1 (i) Since (A,B) is controllable, we can choose K such that $A-BK$ is a stable matrix and let $u_k=-Kx_k$. Then under (A3) ($\{\omega_k\}_{k=0}^{\infty}$ is not necessarily a.s. bounded), it can be verified that $u_k\in\mathcal{U}_{ad}$ (similar to [35]).

(ii) Different from the requirement on the noise in [31–33], we assume that the noise is almost surely bounded in (A3). In fact, (A3) can be alternated by the following: $\{\omega_k,\mathcal{F}_k\}$ is a sequence of d-dimensional independent and identically distributed random vectors satisfying (4) and $E(\omega_k)=\mathbf{0}$ and $\omega_{k,1}$, $\omega_{k,2}$, $\cdots$, $\omega_{k,d}$ are mutually independent for give k and $E(\|\omega_k\|^8)$ exists.

When system coefficients A, B, and C are known, the following result holds according to Theorem 3.5 and Remark 3.4 of [36].

Lemma 2.2 (Theorem 3.5 of [36]) *If* (A2) *holds and* $\{\omega_k,\mathcal{F}_k\}_{k=0}^{\infty}$ *is a martingale differ-*

ence sequence with

$$\begin{cases} \sup\limits_{k\geq 0} E(\|\omega_k\|^2|\mathcal{F}_{k-1}) < \infty \quad a.s., \\ \lim\limits_{k\to\infty} \dfrac{1}{k}\sum\limits_{i=1}^{k} \omega_i\omega_i^{\mathrm{T}} = V > 0 \quad a.s., \end{cases} \tag{5}$$

then the algebraic Riccati equation

$$P^* = A^{\mathrm{T}}P^*A - A^{\mathrm{T}}P^*B(R + B^{\mathrm{T}}P^*B)^{-1}B^{\mathrm{T}}P^*A + Q \tag{6}$$

has a unique positive definite solution P^*, the optimal control minimizing $J(u)$ given by (2) is

$$u_k^* = -(R + B^{\mathrm{T}}P^*B)^{-1}B^{\mathrm{T}}P^*Ax_k := -K^*x_k \tag{7}$$

and the matrix

$$F^* = A - BK^* \tag{8}$$

is stable. The optimal value of the cost function is $J(u^*) = \mathrm{tr}(P^*CVC^{\mathrm{T}})$.

If V is known, then we can obtain the optimal value of the cost function (2). Generally the algebraic Riccati equation (6) is hard to solve and some iteration algorithms were proposed to approximate P^* (see [37, 38]).

VI Algorithm

Step 1 Set $n = 0$, $P_0 = \mathbf{0}$, $K_0 = \mathbf{0}$ and select a sufficiently small constant $\varepsilon > 0$.

Step 2 Compute P_{n+1} by

$$P_{n+1} = A^{\mathrm{T}}P_nA - A^{\mathrm{T}}P_nB(R + B^{\mathrm{T}}P_nB)^{-1}B^{\mathrm{T}}P_nA + Q. \tag{9}$$

Step 3 Update the policy

$$K_{n+1} = (R + B^{\mathrm{T}}P_{n+1}B)^{-1}B^{\mathrm{T}}P_{n+1}A. \tag{10}$$

Step 4 Stop if $\|P_{n+1} - P_n\| < \varepsilon$, and let P_n be an approximation of P^*, otherwise set $n \leftarrow n+1$ and go to Step 2.

In [38], it is proved that the sequences $\{P_n\}_{n=0}^{\infty}$ and $\{K_n\}_{n=0}^{\infty}$ computed from the above VI algorithm can converge to the solution P^* of the algebraic Riccati equation (6) and the optimal control gain K^* respectively.

In the above VI algorithm, we need to know the parameters A and B of the discrete-time linear system. To obtain the optimal value of the cost function (2), V in (4) and the coefficient matrix C of the additive noise are required to be known. However, in some practical problems, the matrices A, B, C, and V are difficult to be accurately obtained. In this work, we will develop a VI-based SADP method to obtain the optimal controller.

3 VI-Based SADP Design for Optimal Control

3.1 System Parameters Transformation

Let $u_i \in \mathcal{U}_{ad}, i = 0, 1, \cdots, k-1$, where k is the running time of the system. For all $P \in \boldsymbol{S}_r$ and $x_i \in \mathbb{R}^r$, by $x_{i+1} = Ax_i + Bu_i + C\omega_{i+1}$, we have

$$\begin{aligned} x_{i+1}^{\mathrm{T}} P x_{i+1} &= (Ax_i + Bu_i + C\omega_{i+1})^{\mathrm{T}} P (Ax_i + Bu_i + C\omega_{i+1}) \\ &= \begin{bmatrix} x_i^{\mathrm{T}} & u_i^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} A^{\mathrm{T}}PA & A^{\mathrm{T}}PB \\ B^{\mathrm{T}}PA & B^{\mathrm{T}}PB \end{bmatrix} \begin{bmatrix} x_i \\ u_i \end{bmatrix} + 2(Ax_i + Bu_i)^{\mathrm{T}} PC\omega_{i+1} \\ &\quad + \omega_{i+1}^{\mathrm{T}} C^{\mathrm{T}} PC\omega_{i+1}. \end{aligned} \tag{11}$$

Define

$$H := \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{bmatrix} A^{\mathrm{T}}PA & A^{\mathrm{T}}PB \\ B^{\mathrm{T}}PA & B^{\mathrm{T}}PB \end{bmatrix}. \tag{12}$$

It follows from (11) and (12) that

$$\begin{aligned} x_{i+1}^{\mathrm{T}} P x_{i+1} &= \begin{bmatrix} x_i^{\mathrm{T}} & u_i^{\mathrm{T}} \end{bmatrix} H \begin{bmatrix} x_i \\ u_i \end{bmatrix} + 2(Ax_i + Bu_i)^{\mathrm{T}} PC\omega_{i+1} + \omega_{i+1}^{\mathrm{T}} C^{\mathrm{T}} PC\omega_{i+1} \\ &= \left[\operatorname{vecv}\left(\begin{bmatrix} x_i \\ u_i \end{bmatrix} \right) \right]^{\mathrm{T}} \operatorname{vecs}(H) + 2(Ax_i + Bu_i)^{\mathrm{T}} PC\omega_{i+1} + \omega_{i+1}^{\mathrm{T}} C^{\mathrm{T}} PC\omega_{i+1}. \end{aligned} \tag{13}$$

Adding $\operatorname{tr}(PCVC^{\mathrm{T}})$ to the right of (13), and subtracting $\operatorname{tr}(PCVC^{\mathrm{T}})$, we get

$$\begin{aligned} x_{i+1}^{\mathrm{T}} P x_{i+1} = &\left[\operatorname{vecv}\left(\begin{bmatrix} x_i \\ u_i \end{bmatrix} \right) \right]^{\mathrm{T}} \operatorname{vecs}(H) + \operatorname{tr}(PCVC^{\mathrm{T}}) + 2(Ax_i + Bu_i)^{\mathrm{T}} PC\omega_{i+1} \\ &+ \omega_{i+1}^{\mathrm{T}} C^{\mathrm{T}} PC\omega_{i+1} - \operatorname{tr}(PCVC^{\mathrm{T}}). \end{aligned} \tag{14}$$

Define

$$\theta(A, B, C, V, P) := \begin{bmatrix} \operatorname{vecs}(H) \\ \operatorname{tr}(PCVC^{\mathrm{T}}) \end{bmatrix}, \quad \theta(A, B, C, V, P) \in \mathbb{R}^{\frac{(r+r_1)(r+r_1+1)}{2}+1}, \tag{15}$$

$$\varphi_i := [\left(\operatorname{vecv}\left([x_i^{\mathrm{T}} \quad u_i^{\mathrm{T}}]^{\mathrm{T}}\right)\right)^{\mathrm{T}} \quad 1]^{\mathrm{T}}, \quad \varphi_i \in \mathbb{R}^{\frac{(r+r_1)(r+r_1+1)}{2}+1}, \tag{16}$$

and

$$\varepsilon_i := 2(Ax_i + Bu_i)^{\mathrm{T}} PC\omega_{i+1} + \omega_{i+1}^{\mathrm{T}} C^{\mathrm{T}} PC\omega_{i+1} - \operatorname{tr}(PCVC^{\mathrm{T}}). \tag{17}$$

Then, we obtain

$$x_{i+1}^{\mathrm{T}} P x_{i+1} = \varphi_i^{\mathrm{T}} \theta(A, B, C, V, P) + \varepsilon_i. \tag{18}$$

We define three transformations $\mathcal{T}_1$, $\mathcal{T}_2$ and $\mathcal{T}_3$, such that

$$\begin{aligned} &H_{11} = A^{\mathrm{T}}PA = \mathcal{T}_1(\theta(A,B,C,V,P)), \quad H_{21} = B^{\mathrm{T}}PA = \mathcal{T}_2(\theta(A,B,C,V,P)),\\ &H_{12} = A^{\mathrm{T}}PB = [\mathcal{T}_2(\theta(A,B,C,V,P))]^{\mathrm{T}}, \quad H_{22} = B^{\mathrm{T}}PB = \mathcal{T}_3(\theta(A,B,C,V,P)). \end{aligned} \tag{19}$$

Now, we need to solve $\theta(A,B,C,V,P)$ from (18). In the following, we use the recursive stochastic least squares method ([36], Chap. 4) to approximately solve $\theta(A,B,C,V,P)$ from (18).

3.2 Estimation of Unknown Term $\theta(A,B,C,V,P)$

Multiplying φ_i on both sides of (18), we have that for any finite running time k,

$$\frac{1}{k}\sum_{i=0}^{k-1}\varphi_i\varphi_i^{\mathrm{T}}\theta(A,B,C,V,P) + \frac{1}{k}\sum_{i=0}^{k-1}\varphi_i\varepsilon_i = \frac{1}{k}\sum_{i=0}^{k-1}(\varphi_i x_{i+1}^{\mathrm{T}}Px_{i+1}). \tag{20}$$

If $\frac{1}{k}\sum_{i=0}^{k-1}\varphi_i\varphi_i^{\mathrm{T}}$ is reversible, then from (20), we can get that

$$\begin{aligned} \theta(A,B,C,V,P) = &\left(\frac{1}{k}\sum_{i=0}^{k-1}\varphi_i\varphi_i^{\mathrm{T}}\right)^{-1}\frac{1}{k}\sum_{i=0}^{k-1}\varphi_i x_{i+1}^{\mathrm{T}}Px_{i+1}\\ &-\left(\frac{1}{k}\sum_{i=0}^{k-1}\varphi_i\varphi_i^{\mathrm{T}}\right)^{-1}\frac{1}{k}\sum_{i=0}^{k-1}\varphi_i\varepsilon_i. \end{aligned} \tag{21}$$

Since ε_i is unknown and $\frac{1}{k}\sum_{i=0}^{k-1}\varphi_i\varphi_i^{\mathrm{T}}$ may be irreversible, $\theta(A,B,C,V,P)$ can not be calculated by (21). We use the recursive algorithm for stochastic least squares estimate ([36], Chap. 4):

$$\begin{aligned} \widehat{\theta}(A,B,C,V,P,k+1) &= \widehat{\theta}(A,B,C,V,P,k) + b_kG_k\varphi_k(x_{k+1}^{\mathrm{T}}Px_{k+1} - \varphi_k^{\mathrm{T}}\widehat{\theta}(A,B,C,V,P,k))\\ &= \widehat{\theta}(A,B,C,V,P,k) + b_kG_k\varphi_k[(\mathrm{vecv}(x_{k+1}))^{\mathrm{T}}\mathrm{vecs}(P)\\ &\quad -\varphi_k^{\mathrm{T}}\widehat{\theta}(A,B,C,V,P,k)] \end{aligned} \tag{22}$$

and

$$\begin{cases} G_{k+1} = G_k - b_kG_k\varphi_k\varphi_k^{\mathrm{T}}G_k,\\ b_k = (1+\varphi_k^{\mathrm{T}}G_k\varphi_k)^{-1}. \end{cases} \tag{23}$$

We set $G_0 = \beta_0 I_{\frac{(r+r_1)(r+r_1+1)}{2}+1}$, $\frac{1}{\mathrm{e}} > \beta_0 > 0$, where e is the natural constant, and set $\widehat{\theta}(A,B,C,V,P,0) = \vec{0}$. For such a selection of $\widehat{\theta}(A,B,C,V,P,0)$ and G_0 we get

$$\begin{aligned} \widehat{\theta}(A,B,C,V,P,k) = &\left(\frac{1}{k}\sum_{i=0}^{k-1}\varphi_i\varphi_i^{\mathrm{T}} + \frac{1}{\beta_0 k}I_{\frac{(r+r_1)(r+r_1+1)}{2}+1}\right)^{-1}\\ &\times\frac{1}{k}\sum_{i=0}^{k-1}\varphi_i x_{i+1}^{\mathrm{T}}Px_{i+1}. \end{aligned} \tag{24}$$

In the following, we prove $\lim_{k\to\infty}\widehat{\theta}(A,B,C,V,P,k) = \theta(A,B,C,V,P)$ a.s..

3.3 Convergence of $\widehat{\theta}(A,B,C,V,P,k)$

Let $x_i := [x_{i,1} \quad x_{i,2} \quad \cdots \quad x_{i,r}]^{\mathrm{T}} \in \mathbb{R}^r$ and $u_i := [u_{i,1} \quad u_{i,2} \quad \cdots \quad u_{i,r_1}]^{\mathrm{T}} \in \mathbb{R}^{r_1}$. For the convergence of $\widehat{\theta}(A,B,C,V,P,k)$, we need a persistent excitation condition:

(A6) There exist constants $k_0 > \frac{(r+r_1)(r+r_1+1)}{2} + 1$ and $\alpha_0 > 0$, such that $\frac{1}{k}\sum_{i=0}^{k-1} \varphi_i\varphi_i^{\mathrm{T}} > \alpha_0 I_{\frac{(r+r_1)(r+r_1+1)}{2}+1}$ for all $k > k_0$.

We can design a controller with stochastic excitation to make the system states satisfy (A6). For example,

$$u_i = \overline{u}_i + e_i := -Kx_i + e_i, \tag{25}$$

where K is chosen such that $A - BK$ is a stable matrix, e_i is the excitation signal. Similar to Chapter 6 of [36], we need some conditions for e_i:

(C1) For $i \geq 0$, $e_i \in \mathcal{F}_i$. $\{e_i\}_{i=0}^{\infty}$ is a sequence of r_1-dimensional independent and identically distributed random vectors, and $\{e_i\}_{i=0}^{\infty}$ is independent of $\{\omega_i\}_{i=0}^{\infty}$ satisfying $E(e_i) = \vec{0}$ and $E(e_ie_i^{\mathrm{T}}) = \lim_{k\to\infty}\frac{1}{k}\sum_{i=0}^{k-1} e_ie_i^{\mathrm{T}} = I_{r_1}$.

(C2) Denote $e_i := [e_{i,1} \quad e_{i,2} \quad \cdots \quad e_{i,r_1}]^{\mathrm{T}}$, and let $e_{i,1}, e_{i,2}, \cdots, e_{i,r_1}$ be mutually independent. $\{e_{i,1}\}_{i=0}^{\infty}$, $\{e_{i,2}\}_{i=0}^{\infty}$, $\cdots$, $\{e_{i,r_1}\}_{i=0}^{\infty}$, $\{\omega_{i,s}\}_{i=0}^{\infty}$ $(s = 1, 2, \cdots, d)$ are mutually independent sequences with $E(e_{i,b}^3) = 0$ for $b = 1, 2, \cdots, r_1$, where $\omega_{i,s}$ is given by $\omega_k := [\omega_{k,1} \quad \omega_{k,2} \quad \cdots \quad \omega_{k,d}]^{\mathrm{T}}$.

(C3) There exists a constant $\alpha > 0$ such that $\sup_i |e_{i,b}| \leq \alpha \quad a.s.$ for $b = 1, 2, \cdots, r_1$.

Remark 3.1 (i) Different from conditions for e_i in Chapter 6 of [36], we require that e_i satisfies (C2).

(ii) The family of σ-algebras $\{\mathcal{F}_i\}$ is rich enough such that both ω_i and e_i are $\mathcal{F}_i$-measurable, otherwise, we need to extend $\mathcal{F}_i$ appropriately ([36]).

(iii) By (C3) and (A3) ($\{\omega_k\}_{k=0}^{\infty}$ is not necessarily a.s. bounded), it can be verified that $u_i \in \mathcal{U}_{ad}$ with u_i defined by (25) (similar to [31]).

The following lemma implies that the control law (25) can make the system states satisfy (A6).

Lemma 3.2 *For the closed loop system* (1) *and* (25) *under* (A1)–(A3) *and* (A5), *there are constants* $\overline{\alpha}_0 > 0$ *and* $\overline{k}_0 > 0$ *such that*

$$\lambda_{\min}\left(\sum_{i=0}^{k-1}\varphi_i\varphi_i^{\mathrm{T}}\right) \geq \overline{\alpha}_0 k, \quad \forall k \geq \overline{k}_0. \tag{26}$$

Proof See Appendix A. ∎

Thus, we have the following convergence result.

Theorem 3.3 *If* (A1)–(A3) *and* (A5)–(A6) *hold, then*

$$\lim_{k\to\infty}\widehat{\theta}(A,B,C,V,P,k) = \theta(A,B,C,V,P) \quad a.s., \tag{27}$$

where $\widehat{\theta}(A,B,C,V,P,k)$ *and* $\theta(A,B,C,V,P)$ *are given by* (22) *and* (21) *respectively.*

Proof See Appendix B. ∎

4 VI-Based SADP Algorithm, Optimal Controller, and Related Analysis

By (22) and mathematical induction, we see that $\widehat{\theta}(A,B,C,V,P,k)$ is linear in P. Hence, we have $\widehat{\theta}(A,B,C,V,P,k)=M_k\mathrm{vecs}(P)$, where $M_k\in\mathbb{R}^{(\frac{(r+r_1)(r+r_1+1)}{2}+1)\times\frac{r(r+1)}{2}}$ and $M_0=\mathbf{0}$. Since (11) is satisfied for any $P\in\boldsymbol{S}_r$, by replacing $\widehat{\theta}(A,B,C,V,P,k)$ in (22) with $M_k\mathrm{vecs}(P)$, we obtain

$$M_{k+1}=M_k+b_kG_k\varphi_k[(\mathrm{vecv}(x_{k+1}))^{\mathrm{T}}-\varphi_k^{\mathrm{T}}M_k]. \tag{28}$$

By the convergence of $\widehat{\theta}(A,B,C,V,P,k)$, we get

$$\lim_{k\to\infty}M_k=M\quad\text{a.s.}, \tag{29}$$

where M satisfies $\theta(A,B,C,V,P)=M\mathrm{vecs}(P)$ for all $P\in\boldsymbol{S}_r$. Let $A=[a_1\ a_2\ \cdots\ a_r]$, $B=[b_1\ b_2\ \cdots\ b_{r_1}]$ and $CVC^{\mathrm{T}}=(h_{s_1j})$, where, for $\overline{b}=1,2,\cdots,r$, $a_{\overline{b}}=[a_{1\overline{b}}\ a_{2\overline{b}}\ \cdots\ a_{r\overline{b}}]^{\mathrm{T}}\in\mathbb{R}^r$ is the $\overline{b}$th column of A and for $\check{j}=1,2,\cdots,r_1$, $b_{\check{j}}=[b_{1\check{j}}\ b_{2\check{j}}\ \cdots\ b_{r\check{j}}]^{\mathrm{T}}\in\mathbb{R}^r$ is the $\check{j}$th column of B.

Define $f(a_{\overline{b}},b_{\check{j}}):=[2a_{1\overline{b}}b_{1\check{j}}\ \ a_{1\overline{b}}b_{2\check{j}}+a_{2\overline{b}}b_{1\check{j}}\ \ \cdots\ \ 2a_{r\overline{b}}b_{r\check{j}}]^{\mathrm{T}}\in\mathbb{R}^{\frac{r(r+1)}{2}}$ for $\overline{b}=1,2,\cdots,r$, $\check{j}=1,2,\cdots,r_1$; $f(a_{\overline{b}},a_{\widehat{b}}):=[2a_{1\overline{b}}a_{1\widehat{b}}\ \ a_{1\overline{b}}a_{2\widehat{b}}+a_{2\overline{b}}a_{1\widehat{b}}\ \ \cdots\ \ 2a_{r\overline{b}}a_{r\widehat{b}}]^{\mathrm{T}}\in\mathbb{R}^{\frac{r(r+1)}{2}}$ for $\overline{b}\neq\widehat{b}$, $\overline{b},\ \widehat{b}=1,2,\cdots,r$; $g(CVC^{\mathrm{T}}):=[h_{11}\ \ h_{12}\ \ \cdots\ \ h_{1r}\ \ h_{22}\ \ h_{23}\ \ \cdots\ \ h_{rr}]^{\mathrm{T}}\in\mathbb{R}^{\frac{r(r+1)}{2}}$, and $f(b_{\check{j}},b_{\widehat{j}}):=[2b_{1\check{j}}b_{1\widehat{j}}\ \ b_{1\check{j}}b_{2\widehat{j}}+b_{2\check{j}}b_{1\widehat{j}}\ \ \cdots\ \ 2b_{r\check{j}}b_{r\widehat{j}}]^{\mathrm{T}}\in\mathbb{R}^{\frac{r(r+1)}{2}}$ for $\check{j}\neq\widehat{j}$, $\check{j},\ \widehat{j}=1,2,\cdots,r_1$. Then $M=[\mathrm{vecv}(a_1)\ \ f(a_1,a_2)\ \ \cdots\ \ f(a_1,b_{r_1})\ \ \mathrm{vecv}(a_2)\ \ \cdots\ \ \mathrm{vecv}(b_{r_1})\ \ g(CVC^{\mathrm{T}})]^{\mathrm{T}}$.

Now, we give our VI-based SADP Algorithm:

Algorithm 1 VI-Based SADP Algorithm

Step 1 Set $k=0,\widehat{P}_0>0$, $M_0=\mathbf{0}$, $\widehat{K}_0=\mathbf{0}$, $x_0=x_0^0=\vec{0}$ and select a sufficiently small constant $\varepsilon>0$.

Step 2 Input $u_k=-Kx_k+e_k$ into the system (1), where K is chosen such that $A-BK$ is a stable matrix and e_k is chosen to satisfy (C1)–(C3).

Step 3 Compute M_{k+1} by (23) and (28).

Step 4 Compute

$$\begin{aligned}
&\widehat{\theta}_1(A,B,C,V,\widehat{P}_k,k+1):=M_{k+1}\mathrm{vecs}(\widehat{P}_k),\\
&\widehat{P}_{k+1}=\mathcal{T}_1(\widehat{\theta}_1(A,B,C,V,\widehat{P}_k,k+1))-\big(\mathcal{T}_2(\widehat{\theta}_1(A,B,C,V,\widehat{P}_k,k+1))\big)^{\mathrm{T}}\\
&\qquad\times(R+\mathcal{T}_3(\widehat{\theta}_1(A,B,C,V,\widehat{P}_k,k+1)))^{+}\mathcal{T}_2(\widehat{\theta}_1(A,B,C,V,\widehat{P}_k,k+1))+Q, \qquad (30)\\
&\widehat{\theta}_2(A,B,C,V,\widehat{P}_{k+1},k+1):=M_{k+1}\mathrm{vecs}(\widehat{P}_{k+1}),\\
&\widehat{K}_{k+1}=(R+\mathcal{T}_3(\widehat{\theta}_2(A,B,C,V,\widehat{P}_{k+1},k+1)))^{+}\mathcal{T}_2(\widehat{\theta}_2(A,B,C,V,\widehat{P}_{k+1},k+1)). \qquad (31)
\end{aligned}$$

Step 5 If $\widehat{P}_{k+1}\not>0$, then $\widehat{P}_{k+1}\leftarrow\widehat{P}_0$, set $k\leftarrow k+1$ and go to Step 2. If $\widehat{P}_{k+1}>0$ and $\|\widehat{P}_{k+1}-\widehat{P}_k\|<\varepsilon$, then let $\widehat{P}_k$ be an approximation of P^*, otherwise set $k\leftarrow k+1$ and go to Step 2.

In Algorithm 1, $\|\widehat{P}_{k+1} - \widehat{P}_k\| < \varepsilon$ is a termination condition, which means that we stop the iteration once $\widehat{P}_k$ changes by only a small amount in an iteration. Without this termination condition, we can obtain two sequences $\{\widehat{P}_k\}_{k=0}^{\infty}$ and $\{\widehat{K}_k\}_{k=0}^{\infty}$ and the following convergence result.

Theorem 4.1 *For Algorithm* 1 *under* (A1)–(A5), *we have* $\lim_{k\to\infty} \widehat{P}_k = P^*$ *a.s. and* $\lim_{k\to\infty} \widehat{K}_k = K^*$ *a.s., where* $\{\widehat{P}_k\}_{k=0}^{\infty}$ *and* $\{\widehat{K}_k\}_{k=0}^{\infty}$ *are given by* (30) *and* (31) *respectively.*

Proof See Appendix C. ∎

For the optimality of the control policy, we have the following result:

Theorem 4.2 *For the system* (1) *and the cost function* (2), *if* (A1)–(A5) *are satisfied, then*

$$\widehat{u}_k^0 = -\widehat{K}_k x_k^0 \tag{32}$$

is the optimal controller, where $\widehat{K}_k$ *is given by* (31) *and* x_k^0 *is the closed-loop solution under the controller* $\widehat{u}_k^0$.

Proof See Appendix D. ∎

Remark 4.3 Since $\lim_{k\to\infty} \widehat{P}_k = P^*$ a.s. and $\lim_{k\to\infty} M_k = M$ a.s., we have $\lim_{k\to\infty} M_k \mathrm{vecs}(\widehat{P}_k) = M\mathrm{vecs}(P^*) = \theta(A, B, C, V, P^*)$. The element in the last row of $\theta(A, B, C, V, P^*)$ is $\mathrm{tr}(P^*CVC^{\mathrm{T}})$. Hence, when C and V are unknown, we can obtain an estimate of the optimal value of the cost function (2) by Algorithm 1.

5 Numerical Simulation

In this section, we present a simulation example to show the effectiveness of our VI-based SADP algorithm and the designed controller.

Consider the two-order stochastic linear discrete-time system

$$x_{k+1} = \begin{bmatrix} 1 & 0.7 \\ -0.2 & 0.3 \end{bmatrix} x_k + \begin{bmatrix} 1 & 1 \\ 0 & 0.8 \end{bmatrix} u_k + \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \omega_{k+1}, \tag{33}$$

where $\omega_k \sim U(\overline{D})$ for $k = 0, 1, \cdots$, $\overline{D} = \{x = (x_1, x_2)^{\mathrm{T}} : -0.5 \le x_i \le 0.5, i = 1, 2\}$ and $\omega_0, \omega_1, \cdots$ are independent and identically distributed. $\omega_{i,1}$ and $\omega_{i,2}$ are mutually independent, and $\omega_{i,s} \sim U(-0.5, 0.5)$ for $s = 1, 2$. The cost function is considered as (2) with $R = Q = I_2$. By VI Algorithm when system parameters A, B, and C are known, we have

$$P^* = \begin{bmatrix} 1.5288 & 0.2201 \\ 0.2201 & 1.1947 \end{bmatrix}, \quad K^* = \begin{bmatrix} 0.4606 & 0.2386 \\ 0.1878 & 0.3125 \end{bmatrix}.$$

If the matrix

$$V = \begin{bmatrix} \frac{1}{12} & 0 \\ 0 & \frac{1}{12} \end{bmatrix}$$

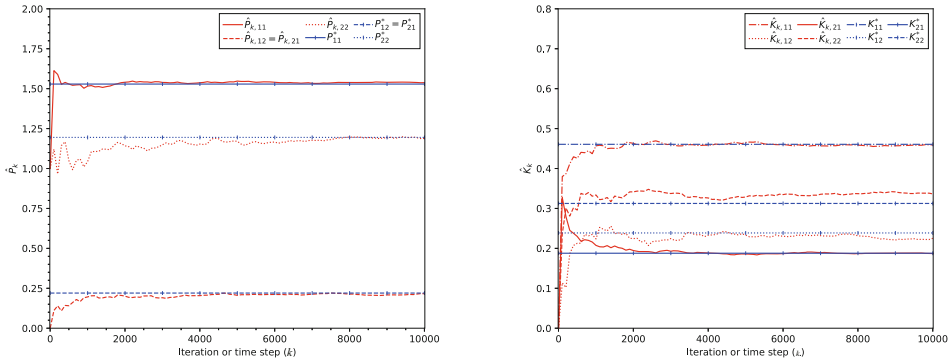
is known, then we get that the optimal value of the cost function is 0.6092.

For our proposed the VI-based SADP Algorithm 1, we take that the initial state is chosen as $x_0 = [0 \ \ 0]^{\mathrm{T}}$. Let $u_k = -Kx_k + e_k$, where

$$K = \begin{bmatrix} 0.5 & 0.7 \\ 0 & 0 \end{bmatrix},$$

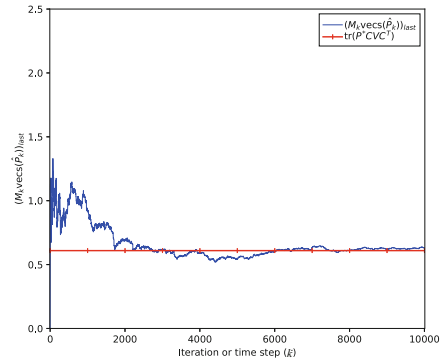
and $e_k \sim U(\widehat{D})$ for $k = 0, 1, \cdots$, $\widehat{D} = \{x = (x_1, x_2)^{\mathrm{T}} : -\sqrt{3} \leq x_i \leq \sqrt{3}, i = 1, 2\}$ and $e_0, e_1, \cdots$ are independent and identically distributed. $e_{i,1}$ and $e_{i,2}$ are mutually independent, and $e_{i,b} \sim U(-\sqrt{3}, \sqrt{3})$ for $b = 1, 2$.

From Figure 1(a) and 1(b), we can see that the elements of $\widehat{P}_k$ and $\widehat{K}_k$ converge to the corresponding optimal values $\widehat{P}^*$ and $\widehat{K}^*$, respectively. In Figure 1(c), $\left(M_k \mathrm{vecs}(\widehat{P}_k)\right)_{last}$ represents the element of the last row of $M_k \mathrm{vecs}(\widehat{P}_k)$, and we get that the estimated value of $\mathrm{tr}(P^* C V C^{\mathrm{T}})$ is 0.6100, which is closely to the optimal value 0.6092. Figure 1(d) shows that the cost function $J(\widehat{u}^0)$ under the controller $\widehat{u}_k^0 = -\widehat{K}_k x_k^0$ can asymptotically approximate the optimal value.

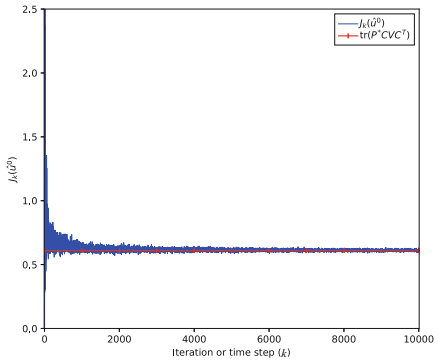


(a) Convergence of $\widehat{P}_k$

(b) Convergence of $\widehat{K}_k$



(c) Convergence of $\left(M_k \mathrm{vecs}(\widehat{P}_k)\right)_{last}$



(d) Convergence of $J_k(\widehat{u}^0)$ under the controller $\widehat{u}_k^0$

Figure 1 Convergence of $\widehat{P}_k$, $\widehat{K}_k$, $\left(M_k \mathrm{vecs}(\widehat{P}_k)\right)_{last}$ in Algorithm 1, and $J_k(\widehat{u}^0)$ under the controller $\widehat{u}_k^0$

Table 1 shows that $\|\widehat{P}_{k+1} - \widehat{P}_k\| < 0.01$ is satisfied when $k > 4800$, $\|\widehat{P}_{k+1} - \widehat{P}_k\| < 0.005$ is satisfied when $k > 7500$, and $\|\widehat{P}_{k+1} - \widehat{P}_k\| < 0.0025$ is satisfied when $k > 9500$.

Table 1 Different values of ε

ε	k_0	$\|\widehat{P}_{k+1} - \widehat{P}_k\| < \varepsilon, k > k_0$
0.01	4800	Yes
0.005	7500	Yes
0.0025	9500	Yes

6 Concluding Remarks

In this paper, we developed a VI-based SADP algorithm to solve the optimal control problem of unknown linear discrete-time systems with additive noise. Our proposed algorithm can directly use the value of system states to obtain the optimal controllers and the optimal cost function without estimating expectation (conditional expectation) and variance (conditional variance) of states or other variables. It should be pointed out that for unknown linear systems with additive noise, the analysis of the optimal control is hard. In this paper, we made a try to give an SADP algorithm for the optimal control of such systems and supply rigorous convergence analysis of the algorithm and the optimality analysis of the controller. In our future work, we will investigate the method of stochastic adaptive dynamic programming for more general cases.

References

[1] Lewis F L, Vrabie D L, and Syrmos V L, *Optimal Control*, John Wiley & Sons Inc., Hoboken, 2012.

[2] Guo J, Zhang J F, and Zhao Y L, Adaptive tracking of a class of first-order systems with binary-valued observations and fixed thresholds, *Journal of Systems Science and Complexity*, 2012, **25**(6): 1041–1051.

[3] Jiang Y and Jiang Z P, A robust adaptive dynamic programming principle for sensorimotor control with signal-dependent noise, *Journal of Systems Science and Complexity*, 2015, **28**(2): 261–288.

[4] Chen H F, Noisy observation based stabilization and optimization for unknown systems, *Journal of Systems Science and Complexity*, 2003, **16**(3): 315–326.

[5] Tang Q Y and Chen H F, Optimal adaptive control with constraint for ARMAX model, *Journal of Systems Science and Complexity*, 1991, **4**(3): 254–263.

[6] Li X X, Peng Z H, Jiao L, et al., Online adaptive Q-learning method for fully cooperative linear quadratic dynamic games, *Science China Information Sciences*, 2019, **62**(12): 1–14.

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

[8] Lewis F L and Vamvoudakis K G, Reinforcement learning for partially observable dynamic processes: Adaptive dynamic programming using measured output data, *IEEE Transactions on Systems, Man, and Cybernetics, Part B* (*Cybernetics*), 2010, **41**(1): 14–25.

[9] Rizvi S A A and Lin Z L, Output feedback Q-learning control for the discrete-time linear quadratic regulator problem, *IEEE Transactions on Neural Networks and Learning Systems*, 2018, **30**(5): 1523–1536.

[10] Kiumarsi B, Lewis F L, Modares H, et al., Reinforcement Q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics, *Automatica*, 2014, **50**(4): 1167–1175.

[11] Jiang Y, Fan J L, Chai T Y, et al., Tracking control for linear discrete-time networked control systems with unknown dynamics and dropout, *IEEE Transactions on Neural Networks and Learning Systems*, 2017, **29**(10): 4607–4620.

[12] He P and Jagannathan S, Reinforcement learning-based output feedback control of nonlinear systems with input constraints, *IEEE Transactions on Systems, Man, and Cybernetics, Part B* (*Cybernetics*), 2005, **35**(1): 150–154.

[13] Wei Q L and Liu D R, A novel policy iteration based deterministic Q-learning for discrete-time nonlinear systems, *Science China Information Sciences*, 2015, **58**(12): 1–15.

[14] Wang D, Liu D R, Li H L, et al., An approximate optimal control approach for robust stabilization of a class of discrete-time nonlinear systems with uncertainties, *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2015, **46**(5): 713–717.

[15] Liu R R, Li Y, and Liu X K, Linear-quadratic optimal control for unknown mean-field stochastic discrete-time system via adaptive dynamic programming approach, *Neurocomputing*, 2018, **282**: 16–24.

[16] Liu X K, Liu R R, and Li Y, Infinite time linear quadratic stackelberg game problem for unknown stochastic discrete-time systems via adaptive dynamic programming approach, *Asian Journal of Control*, 2021, **23**(2): 937–948.

[17] Gravell B, Ganapathy K, and Summers T, Policy iteration for linear quadratic games with stochastic parameters, *IEEE Control Systems Letters*, 2020, **5**(1): 307–312.

[18] Wang J S and Yang G H, Output-feedback control of unknown linear discrete-time systems with stochastic measurement and process noise via approximate dynamic programming, *IEEE Transactions on Cybernetics*, 2017, **48**(7): 1977–1988.

[19] Han K Z, Feng J, and Yao Y, An integrated data-driven Markov parameters sequence identification and adaptive dynamic programming method to design fault-tolerant optimal tracking control for completely unknown model systems, *Journal of the Franklin Institute*, 2017, **354**(13): 5280–5301.

[20] Wong W C and Lee J H, A reinforcement learning-based scheme for direct adaptive optimal control of linear stochastic systems, *Optimal Control Applications and Methods*, 2010, **31**(4): 365–374.

[21] Yaghmaie F A and Gustafsson F, Using reinforcement learning for model-free linear quadratic control with process and measurement noises, *Proceedings of the* 58*th IEEE Conference on Decision and Control* (*CDC*), Nice, France, Dec. 11–13, 2019, 6510–6517.

[22] Abbasi-Yadkori Y, Lazić N, and Szepesvári C, Model-free linear quadratic control via reduction to expert prediction, *Proceedings of the* 22*nd International Conference on Artificial Intelligence*

and Statistics (*AISTATS*), Okinawa, Japan, Apr. 16–18, 2019, 3108–3117.

[23] Xu X, Chen H, Lian C Q, et al., Learning-based predictive control for discrete-time nonlinear systems with stochastic disturbances, *IEEE Transactions on Neural Networks and Learning Systems*, 2018, **29**(12): 6202–6213.

[24] Liang M M, Wang D, and Liu D R, Neuro-optimal control for discrete stochastic processes via a novel policy iteration algorithm, *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2019, **50**(11): 3972–3985.

[25] Liang M M, Wang D, and Liu D R, Improved value iteration for neural-network-based stochastic optimal control design, *Neural Networks*, 2020, **124**: 280–295.

[26] M'sahli F, Fayeche C, Abdennour R B, et al., Application of adaptive controllers for the temperature control of a semi-batch reactor, *International Journal of Computational Engineering Science*, 2001, **2**(2): 287–307.

[27] Haas S M, Frei M G, Osorio I, et al., EEG ocular artifact removal through ARMAX model system identification using extended least squares, *Communications in Information and Systems*, 2003, **3**(1): 19–40.

[28] Deisenroth M P, Fox D, and Rasmussen C E, Gaussian processes for data-efficient learning in robotics and control, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, **37**(2): 408–423.

[29] Sethi S P, Suo W, Taksar M I, et al., Optimal production planning in a multi-product stochastic manufacturing system with long-run average cost, *Discrete Event Dynamic Systems*, 1998, **8**(1): 37–54.

[30] Borkar V S, Ergodic control of diffusion processes, *Proceedings of the International Congress of Mathematicians* (*ICM*), Madrid, Spain, 2006: 1299–1309.

[31] Chen H F and Guo L, Optimal stochastic adaptive control with quadratic index, *International Journal of Control*, 1986, **43**(3): 869–881.

[32] Chen H F and Guo L, Stochastic adaptive control for a general quadratic cost, *Journal of Systems Science and Mathematical Sciences*, 1987, **7**(4): 289–302.

[33] Guo L, Self-convergence of weighted least-squares with applications to stochastic adaptive control, *IEEE Transactions on Automatic Control*, 1996, **41**(1): 79–89.

[34] Sutton R S, Barto A G, and Williams R J, Reinforcement learning is direct adaptive optimal control, *IEEE Control Systems Magazine*, 1992, **12**(2): 19–22.

[35] Ma C Q, Li T, and Zhang J F, Linear quadratic decentralized dynamic games for large population discrete-time stochastic multi-agent systems, *Journal of Systems Science and Mathematical Sciences*, 2007, **27**(3): 464–480.

[36] Chen H F and Guo L, *Identification and Stochastic Adaptive Control*, Springer Science & Business Media, New York, 1991.

[37] Gao W N, Jiang Y, Jiang Z P, et al., Output-feedback adaptive optimal control of interconnected systems based on robust adaptive dynamic programming, *Automatica*, 2016, **72**: 37–45.

[38] Lancaster P and Rodman L, *Algebraic Riccati Equations*, Oxford University Press Inc., New York, 1995.

Appendix A: Proof of Lemma 3.2

Firstly, we show that $\{u_i\}_{i=0}^{\infty}$ and $\{x_i\}_{i=0}^{\infty}$ are a.s. bounded.

By (25) and $e_i \in \mathcal{F}_i$, we get $u_i \in \mathcal{F}_i$. For (1) and $u_{i-1} = -Kx_{i-1} + e_{i-1}$, we have

$$x_i = (A - BK)^i x_0 + \sum_{l=0}^{i-1}(A - BK)^l B e_{i-l-1} + \sum_{l=0}^{i-1}(A - BK)^l C\omega_{i-l}. \tag{34}$$

Since $A-BK$ is the stable matrix, there are $S_0 > 0$ and $\overline{\lambda} \in (0,1)$ such that $\|A-BK\|^i \le S_0\overline{\lambda}^i$ for all $i \ge 0$ by [35].

It follows from a.s. boundedness of $\{\omega_k\}_{k=0}^{\infty}$, $\|A-BK\|^i \le S_0\overline{\lambda}^i$, and (C3) that there exists a constant $\overline{\alpha} > 0$ such that $\sup_i |x_{i,\overline{b}}| < \overline{\alpha}$ a.s. for $\overline{b} = 1, 2, \cdots, r$. This together with (25) and (C3) implies that there exists a constant $\widetilde{\alpha} > 0$ such that $\sup_i |u_{i,b}| < \widetilde{\alpha}$ a.s. for $b = 1, 2, \cdots, r_1$.

Secondly, we prove that (26) is true by the proof idea of Theorem 6.2 in [36].

Since (26) is equivalent to

$$\liminf_{k\to\infty} k^{-1}\lambda_{\min}\Big(\sum_{i=0}^{k-1}\varphi_i\varphi_i^{\mathrm{T}}\Big) > 0, \tag{35}$$

we will prove that (35) holds in the following.

Now, (1) can be written in the following form $(I - Az)x_{k+1} = Bzu_{k+1} + C\omega_{k+1}$, where z denotes the shift-back operator: $zx_{k+1} = x_k$. $\det(I - Az)$ denotes the determinant of $I - Az$. Let

$$g_k := (\det(I - Az))^2\varphi_k, \tag{36}$$

$$\det(I - Az) = \overline{a}_0 + \overline{a}_1 z + \cdots + \overline{a}_p z^p, \quad p \le r, \tag{37}$$

where $\overline{a}_0, \overline{a}_1, \cdots, \overline{a}_p$ depend on the elements of A.

By the Cauchy-Schwarz inequality, we have

$$\lambda_{\min}\Big(\sum_{i=0}^{k-1} g_i g_i^{\mathrm{T}}\Big) = \inf_{\|y\|=1}\sum_{i=0}^{k-1}(y^{\mathrm{T}} g_i g_i^{\mathrm{T}} y) \le (p+1)^2\Big(\sum_{j=0}^{p}\overline{a}_j^2\Big)^2\lambda_{\min}\Big(\sum_{i=0}^{k-1}\varphi_i\varphi_i^{\mathrm{T}}\Big). \tag{38}$$

Thus, for (35) it suffices to show that

$$\liminf_{k\to\infty} k^{-1}\lambda_{\min}\Big(\sum_{i=0}^{k-1} g_i g_i^{\mathrm{T}}\Big) > 0. \tag{39}$$

Now we use proof by contradiction to prove (39).

If (39) were not true, then there would exist a vector sequence $\{\beta_{k_m}\}$:

$$\beta_{k_m} = [\alpha_{k_m,1} \ \ \alpha_{k_m,2} \ \ \cdots \ \ \alpha_{k_m,\frac{(r+r_1)(r+r_1+1)}{2}+1}]^{\mathrm{T}} \in \mathbb{R}^{\frac{(r+r_1)(r+r_1+1)}{2}+1},$$

such that $\|\beta_{k_m}\| = 1$ and

$$\lim_{m\to\infty} k_m^{-1}\Big(\sum_{i=0}^{k_m-1}(\beta_{k_m}^{\mathrm{T}} g_i)^2\Big) = 0. \tag{40}$$

$\mathrm{Adj}(I-Az)$ denotes the adjoint matrix of $I-Az$. Let

$$(\mathrm{Adj}(I-Az))^{\mathrm{T}}=[a_1(z)\ a_2(z)\ \cdots\ a_r(z)], \tag{41}$$

$$[(\det(I-Az))I_{r_1}\ \ 0_{r_1\times d}]^{\mathrm{T}}=[\det(I-Az)\gamma_1\ \det(I-Az)\gamma_2\ \cdots\ \det(I-Az)\gamma_{r_1}], \tag{42}$$

where, for $\overline{b}=1,2,\cdots,r$, $a_{\overline{b}}(z)\in\mathbb{R}^r$ is the $\overline{b}$th column of $(\mathrm{Adj}(I-Az))^{\mathrm{T}}$, and for $b=1,2,\cdots,r_1$, $\gamma_b\in\mathbb{R}^{r_1+d}$ is the column vector of the bth row with 1 and other rows with 0. Notice that

$$(\det(I-Az))x_k=(\mathrm{Adj}(I-Az))[Bz\ \ C]\begin{bmatrix}u_k\\ \omega_k\end{bmatrix}, \tag{43}$$

$$(\det(I-Az))u_k=[(\det(I-Az))I_{r_1}\ \ 0]\begin{bmatrix}u_k\\ \omega_k\end{bmatrix}, \tag{44}$$

and definition of φ_i. Then we have

$$\begin{aligned}\beta_{k_m}^{\mathrm{T}}g_i=&\left[\alpha_{k_m,1}\left(\mathrm{vec}\left(\begin{bmatrix}B^{\mathrm{T}}z\\ C^{\mathrm{T}}\end{bmatrix}a_1(z)a_1^{\mathrm{T}}(z)[Bz\ \ C]\right)\right)^{\mathrm{T}}+\alpha_{k_m,2}\left(\mathrm{vec}\left(\begin{bmatrix}B^{\mathrm{T}}z\\ C^{\mathrm{T}}\end{bmatrix}a_1(z)a_2^{\mathrm{T}}(z)\right.\right.\\ &\left.\left.\times[Bz\ \ C]\right)\right)^{\mathrm{T}}+\cdots+\alpha_{k_m,\frac{(r+r_1)(r+r_1+1)}{2}}(\det(I-Az))^2(\mathrm{vec}(\gamma_{r_1}\gamma_{r_1}^{\mathrm{T}}))^{\mathrm{T}}\Bigg]\\ &\times\left(\begin{bmatrix}u_i\\ \omega_i\end{bmatrix}\otimes\begin{bmatrix}u_i\\ \omega_i\end{bmatrix}\right)+\alpha_{k_m,\frac{(r+r_1)(r+r_1+1)}{2}+1}\big(\det(I-Az)\big)^2.\end{aligned} \tag{45}$$

Let

$$\begin{aligned}L_{k_m}(z)=&\alpha_{k_m,1}\left(\mathrm{vec}\left(\begin{bmatrix}B^{\mathrm{T}}z\\ C^{\mathrm{T}}\end{bmatrix}a_1(z)a_1^{\mathrm{T}}(z)[Bz\ \ C]\right)\right)^{\mathrm{T}}+\alpha_{k_m,2}\left(\mathrm{vec}\left(\begin{bmatrix}B^{\mathrm{T}}z\\ C^{\mathrm{T}}\end{bmatrix}a_1(z)a_2^{\mathrm{T}}(z)\right.\right.\\ &\left.\left.\times[Bz\ \ C]\right)\right)^{\mathrm{T}}+\cdots+\alpha_{k_m,\frac{(r+r_1)(r+r_1+1)}{2}}\big(\det(I-Az)\big)^2(\mathrm{vec}(\gamma_{r_1}\gamma_{r_1}^{\mathrm{T}}))^{\mathrm{T}}\\ =&\sum_{j=0}^{2p}[\rho_{k_m,1}^{(j)}\ \rho_{k_m,2}^{(j)}\ \cdots\ \rho_{k_m,(r_1+d)^2}^{(j)}]z^j,\end{aligned} \tag{46}$$

where, for $\tau=1,2,\cdots,(r_1+d)^2$, $\rho_{k_m,\tau}^{(j)}\in\mathbb{R}$. By $\|\beta_{k_m}\|=1$, we have that $\rho_{k_m,1}^{(j)},\rho_{k_m,2}^{(j)},\cdots,\rho_{k_m,(r_1+d)^2}^{(j)}$ are bounded. By (45) and (46), we get

$$\begin{aligned}\lim_{m\to\infty}k_m^{-1}\sum_{i=0}^{k_m-1}(\beta_{k_m}^{\mathrm{T}}g_i)^2=&\lim_{m\to\infty}k_m^{-1}\sum_{i=0}^{k_m-1}\Big[\rho_{k_m,1}^{(0)}u_{i,1}^2+(\rho_{k_m,2}^{(0)}+\rho_{k_m,r_1+d+1}^{(0)})u_{i,1}u_{i,2}\\ &+\cdots+\rho_{k_m,(r_1+d)^2}^{(2p)}\omega_{i-p,d}^2+\alpha_{k_m,\frac{(r+r_1)(r+r_1+1)}{2}+1}\\ &\times\big(\det(I-Az)\big)^2\Big]^2\\ =&\,0.\end{aligned} \tag{47}$$

By (25), (47) and combining terms containing $e_{i,1}$, we have

$$\begin{aligned}\lim_{m\to\infty} k_m^{-1}\sum_{i=0}^{k_m-1}(\beta_{k_m}^{\mathrm{T}}g_i)^2 = \lim_{m\to\infty} k_m^{-1}\sum_{i=0}^{k_m-1}\Big[\Big(2\rho_{k_m,1}^{(0)}\overline{u}_{i,1}+\Big(\rho_{k_m,2}^{(0)}+\rho_{k_m,r_1+d+1}^{(0)}\Big)u_{i,2}\\ +\cdots+\Big(\gamma_{k_m,r_1+d}^{(p,0)}+\gamma_{k_m,(r_1+d-1)(r_1+d)+1}^{(p,0)}\Big)\omega_{i-p,d}\Big)e_{i,1}\\ +\rho_{k_m,1}^{(0)}\overline{u}_{i,1}^2+\rho_{k_m,1}^{(0)}e_{i,1}^2+\cdots+\rho_{k_m,(r_1+d)^2}^{(2p)}\omega_{i-p,d}^2\\ +\alpha_{k_m,\frac{(r+r_1)(r+r_1+1)}{2}+1}\big(\det(I-Az)\big)^2\Big]^2\\ =0,\end{aligned}\tag{48}$$

where $\gamma_{k_m,r_1+d}^{(p,0)}\in\mathbb{R}$, and the sum of the coefficients of $u_{i-1,1}\omega_{i-(p-1),d}$, $u_{i-2,1}\omega_{i-(p-2),d}$, $\cdots$, $u_{i-p,1}\omega_{i,d}$ and $\gamma_{k_m,r_1+d}^{(p,0)}$ is $\rho_{k_m,r_1+d}^{(p)}$. For $\gamma_{k_m,(r_1+d-1)(r_1+d)+1}^{(p,0)}$, $\rho_{k_m,1}^{(\breve{i})}$, $\rho_{k_m,2}^{(\breve{i})}$, $\cdots$, $\rho_{k_m,(r_1+d)^2}^{(\breve{i})}$, $\breve{i}=1,2,\cdots,2p-1$, we have the same argument. The coefficients of all terms in (48) are bounded by $\|\beta_{k_m}\|=1$.

Let $\mathcal{G}_i^{(1)}=\sigma\{e_{j,1},e_{j+1,2},\cdots,e_{j+1,r_1},\omega_{j+1,1},\omega_{j+1,2},\cdots,\omega_{j+1,d},0\le j\le i\}$. It follows from (C1) and (C2) that $\{e_{i,1},\mathcal{G}_i^{(1)}\}$ and $\{e_{i,1}^3,\mathcal{G}_i^{(1)}\}$ are martingale difference sequences. Notice the facts that $x_{i,\overline{b}}$, $\overline{u}_{i,b}$, $e_{i,b}$, $u_{i,b}$, and $\omega_{i,s}$ are a.s. bounded for $\overline{b}=1,2,\cdots,r$, $b=1,2,\cdots,r_1$, and $s=1,2,\cdots,d$ and $\rho_{k_m,1}^{(j)}$, $\rho_{k_m,2}^{(j)}$, $\cdots$, $\rho_{k_m,(r_1+d)^2}^{(j)}$, $\alpha_{k_m,\frac{(r+r_1)(r+r_1+1)}{2}+1}\big(\det(I-Az)\big)^2$, and β_{k_m} are bounded for $j=0,1,\cdots,2p$. Then, by Theorem 2.8 of [36] we have that for any $\eta>0$,

$$\begin{aligned}&k_m^{-1}\sum_{i=0}^{k_m-1}\Big[\rho_{k_m,1}^{(0)}\overline{u}_{i,1}^2+\rho_{k_m,1}^{(0)}e_{i,1}^2+\cdots+\rho_{k_m,(r_1+d)^2}^{(2p)}\omega_{i-p,d}^2+\alpha_{k_m,\frac{(r+r_1)(r+r_1+1)}{2}+1}\\ &\times\big(\det(I-Az)\big)^2\Big]\Big[2\rho_{k_m,1}^{(0)}\overline{u}_{i,1}+(\rho_{k_m,2}^{(0)}+\rho_{k_m,r_1+d+1}^{(0)})u_{i,2}\\ &+\cdots+(\gamma_{k_m,r_1+d}^{(p,0)}+\gamma_{k_m,(r_1+d-1)(r_1+d)+1}^{(p,0)})\omega_{i-p,d}\Big]e_{i,1}\\ &=O\Big(k_m^{-1}\times k_m^{\frac{1}{2}}\log^{\frac{1}{2}+\eta}(k_m+\mathrm{e})\Big)\to 0\quad(m\to\infty)\ \text{ a.s..}\end{aligned}\tag{49}$$

This together with (48) implies

$$\begin{aligned}\lim_{m\to\infty} k_m^{-1}\sum_{i=0}^{k_m-1}\big[\big(2\rho_{k_m,1}^{(0)}\overline{u}_{i,1}+(\rho_{k_m,2}^{(0)}+\rho_{k_m,r_1+d+1}^{(0)})u_{i,2}\\ +\cdots+(\gamma_{k_m,r_1+d}^{(p,0)}+\gamma_{k_m,(r_1+d-1)(r_1+d)+1}^{(p,0)})\omega_{i-p,d}\big)e_{i,1}\big]^2\\ =0\end{aligned}\tag{50}$$

and

$$\begin{aligned}\lim_{m\to\infty} k_m^{-1}\sum_{i=0}^{k_m-1}\big(\rho_{k_m,1}^{(0)}\overline{u}_{i,1}^2+\rho_{k_m,1}^{(0)}e_{i,1}^2+\cdots+\rho_{k_m,(r_1+d)^2}^{(2p)}\omega_{i-p,d}^2\\ +\alpha_{k_m,\frac{(r+r_1)(r+r_1+1)}{2}+1}\big(\det(I-Az)\big)^2\big)^2\\ =0.\end{aligned}\tag{51}$$

It follows from (C1) and (C2) that $\{e_{i,1}^2-1,\mathcal{G}_i^{(1)}\}$ is the martingale difference sequence. Notice the facts that $\overline{u}_{i,b}$, $u_{i,b}$, and $\omega_{i,s}$ are a.s. bounded for $b=1,2,\cdots,r_1$ and $s=1,2,\cdots,d$ and $\rho_{k_m,1}^{(l_1)}$, $\rho_{k_m,2}^{(l_1)}$, $\cdots$, $\rho_{k_m,(r_1+d)^2}^{(l_1)}$, and β_{k_m} are bounded for $l_1=0,1,\cdots,p$. Then, by Theorem 2.8 of [36], we obtain

$$\begin{aligned}&\lim_{m\to\infty}k_m^{-1}\sum_{i=0}^{k_m-1}\big(2\rho_{k_m,1}^{(0)}\overline{u}_{i,1}+(\rho_{k_m,2}^{(0)}+\rho_{k_m,r_1+d+1}^{(0)})u_{i,2}\\&+\cdots+(\gamma_{k_m,r_1+d}^{(p,0)}+\gamma_{k_m,(r_1+d-1)(r_1+d)+1}^{(p,0)})\omega_{i-p,d}\big)^2(e_{i,1}^2-1)\\&=0,\end{aligned}\tag{52}$$

and by (50)

$$\begin{aligned}&\lim_{m\to\infty}k_m^{-1}\sum_{i=0}^{k_m-1}\big(2\rho_{k_m,1}^{(0)}\overline{u}_{i,1}+(\rho_{k_m,2}^{(0)}+\rho_{k_m,r_1+d+1}^{(0)})u_{i,2}\\&+\cdots+(\gamma_{k_m,r_1+d}^{(p,0)}+\gamma_{k_m,(r_1+d-1)(r_1+d)+1}^{(p,0)})\omega_{i-p,d}\big)^2\\&=0.\end{aligned}\tag{53}$$

Notice that $\lim_{k\to\infty}\frac{1}{k}\sum_{i=0}^{k-1}e_ie_i^{\mathrm{T}}=I_{r_1}$. Thus continuing similar argument for (53), we get

$$\lim_{m\to\infty}(\rho_{k_m,2}^{(0)}+\rho_{k_m,r_1+d+1}^{(0)})=0.\tag{54}$$

Similarly,

$$\begin{cases}\lim\limits_{m\to\infty}\rho_{k_m,i_1(r_1+d)+i_1+1}^{(j)}=0,\quad i_1=0,1,\cdots,r_1-1;\\ \lim\limits_{m\to\infty}\rho_{k_m,(r_1+j_1)(r_1+d)+r_1+j_1+1}^{(j)}=0,\quad j_1=0,1,\cdots,d-1;\end{cases}\tag{55}$$

$$\begin{aligned}&\lim_{m\to\infty}\big(\rho_{k_m,(s_1-1)(r_1+d)+i_2}^{(j)}+\rho_{k_m,(i_2-1)(r_1+d)+s_1}^{(j)}\big)=0,\\&s_1=1,2,\cdots,r_1;\quad i_2=s_1+1,s_1+2,\cdots,r_1+d;\end{aligned}\tag{56}$$

$$\begin{aligned}&\lim_{m\to\infty}\big(\rho_{k_m,(r_1+s_2-1)(r_1+d)+i_2+r_1}^{(j)}+\rho_{k_m,(r_1+i_2-1)(r_1+d)+s_2+r_1}^{(j)}\big)=0,\\&s_2=1,2,\cdots,d-1;\quad i_2=s_2+1,s_2+2,\cdots,d;\end{aligned}\tag{57}$$

$$\lim_{m\to\infty}\alpha_{k_m,\frac{(r+r_1)(r+r_1+1)}{2}+1}=0,\tag{58}$$

where $j=0,1,\cdots,2p$.

Since $\{\alpha_{k_m,\frac{(r+r_1)(r+r_1+1)}{2}+1}\}$ is bounded and (58), there is a subsequence $\{\alpha_{k_{m_{\overline{\tau}}},\frac{(r+r_1)(r+r_1+1)}{2}+1}\}$ satisfying

$$\lim_{\overline{\tau}\to\infty}\alpha_{k_{m_{\overline{\tau}}},\frac{(r+r_1)(r+r_1+1)}{2}+1}=0.\tag{59}$$

Since $\{\beta_{k_m}\}$ is bounded, there exists a convergent subsequence tending to a limit

$$\beta=[\alpha_1\ \ \alpha_2\ \ \cdots\ \ \alpha_{\frac{(r+r_1)(r+r_1+1)}{2}}\ \ 0]^{\mathrm{T}}$$

with unit norm.

Let

$$H_{k_m}(z)=\alpha_{k_m,1}\begin{bmatrix}B^{\mathrm{T}}z\\ C^{\mathrm{T}}\end{bmatrix}a_1(z)a_1^{\mathrm{T}}(z)[Bz\ \ C]+\alpha_{k_m,2}\begin{bmatrix}B^{\mathrm{T}}z\\ C^{\mathrm{T}}\end{bmatrix}a_1(z)a_2^{\mathrm{T}}(z)[Bz\ \ C]$$
$$+\cdots+\alpha_{k_m,\frac{(r+r_1)(r+r_1+1)}{2}}\big(\det(I-Az)\big)^2\gamma_{r_1}\gamma_{r_1}^{\mathrm{T}}. \tag{60}$$

For $j=0,1,\cdots,2p$; $\rho^{(j)}_{k_m,1}$, $\rho^{(j)}_{k_m,2}$, $\cdots$, $\rho^{(j)}_{k_m,(r_1+d)^2}$ are bounded. This together with (46) and (55)–(57) implies that there is a subsequence $\{H_{k_{m_{\overline{\tau}}}}(z)\}$ satisfying

$$\lim_{\overline{\tau}\to\infty}H_{k_{m_{\overline{\tau}}}}(z)=H(z), \tag{61}$$

where

$$H(z)=\alpha_1\begin{bmatrix}B^{\mathrm{T}}z\\ C^{\mathrm{T}}\end{bmatrix}a_1(z)a_1^{\mathrm{T}}(z)[Bz\ \ C]+\alpha_2\begin{bmatrix}B^{\mathrm{T}}z\\ C^{\mathrm{T}}\end{bmatrix}a_1(z)a_2^{\mathrm{T}}(z)[Bz\ \ C]$$
$$+\cdots+\alpha_{\frac{(r+r_1)(r+r_1+1)}{2}}\big(\det(I-Az)\big)^2\gamma_{r_1}\gamma_{r_1}^{\mathrm{T}} \tag{62}$$

and $H(z)$ is an anti-symmetric matrix.

Since $H(z)$ is the anti-symmetric matrix, we have

$$H(z)+H(z)^{\mathrm{T}}=\mathbf{0}. \tag{63}$$

By (A5) and comparing coefficients of elements on both sides of (63), we obtain

$$\alpha_1=\alpha_2=\cdots=\alpha_{\frac{(r+r_1)(r+r_1+1)}{2}}=0. \tag{64}$$

Thus, it follows from (64) that $\beta=\vec{0}$. However, this is a contradiction with $\|\beta\|=1$. Thus (39) is true. The proof is completed. ∎

Appendix B: Proof of Theorem 3.3

Let $u_i=-Kx_i+e_i$, where K is chosen such that $A-BK$ is a stable matrix, and e_i is chosen to satisfy (C1)–(C3). Thus, by (A3) ($\{\omega_k\}_{k=0}^{\infty}$ is not necessarily a.s. bounded) we have $\{u_i\}\in\mathcal{U}_{ad}$, i.e.,

$$\sum_{i=0}^{k-1}(\|u_i\|^2+\|x_i\|^2)=O(k)\quad\text{a.s..} \tag{65}$$

By Theorem 2.8 of [36] and (A3) ($\{\omega_k\}_{k=0}^{\infty}$ is not necessarily a.s. bounded), we obtain that for any $\delta>0$,

$$\begin{aligned}&\frac{1}{k}\sum_{i=0}^{k-1}(Ax_i+Bu_i)^{\mathrm{T}}PC\omega_{i+1}\\ &=O\bigg[\frac{1}{k}\bigg(\sum_{i=0}^{k-1}(\|x_i\|^2+\|u_i\|^2)\bigg)^{\frac{1}{2}}\bigg(\log\bigg(\sum_{i=0}^{k-1}(\|x_i\|^2+\|u_i\|^2)+\mathrm{e}\bigg)\bigg)^{\frac{1}{2}+\delta}\bigg]\\ &=O\bigg(\frac{1}{\sqrt{k}}\big(\log(k+\mathrm{e})\big)^{\frac{1}{2}+\delta}\bigg)\to 0\quad(k\to\infty)\quad\text{a.s..}\end{aligned} \tag{66}$$

It follows from (4), (17), and (66) that $\lim_{k\to\infty}\frac{1}{k}\sum_{i=0}^{k-1}\varepsilon_i=0$ a.s.. By the proof of Lemma 3.2 and the definition of φ_i, we know that there exists a constant $M^0>0$ such that $\sup_i|\varphi_{i,h}|\le M^0$ a.s. for $h=1,2,\cdots,\frac{(r+r_1)(r+r_1+1)}{2}+1$.
By (21), (24), and Lemma 3.2, we get

$$\begin{aligned}&\lim_{k\to\infty}\big(\widehat{\theta}(A,B,C,V,P,k)-\theta(A,B,C,V,P)\big)\\&=\lim_{k\to\infty}\Bigg\{\Bigg[\bigg(\frac{1}{k}\sum_{i=0}^{k-1}\varphi_i\varphi_i^{\mathrm{T}}+\frac{1}{k\beta_0}I_{\frac{(r+r_1)(r+r_1+1)}{2}+1}\bigg)^{-1}-\bigg(\frac{1}{k}\sum_{i=0}^{k-1}\varphi_i\varphi_i^{\mathrm{T}}\bigg)^{-1}\Bigg]\\&\quad\times\frac{1}{k}\sum_{i=0}^{k-1}\varphi_i x_{i+1}^{\mathrm{T}}Px_{i+1}\Bigg\}+\lim_{k\to\infty}\Bigg\{\bigg(\frac{1}{k}\sum_{i=0}^{k-1}\varphi_i\varphi_i^{\mathrm{T}}\bigg)^{-1}\frac{1}{k}\sum_{i=0}^{k-1}\varphi_i\varepsilon_i\Bigg\},\quad k\ge\overline{k}_0.\end{aligned}\tag{67}$$

Let

$$\begin{cases}E_k=\bigg(\frac{1}{k}\sum_{i=0}^{k-1}\varphi_i\varphi_i^{\mathrm{T}}+\frac{1}{k\beta_0}I_{\frac{(r+r_1)(r+r_1+1)}{2}+1}\bigg)^{-1}, & k\ge\overline{k}_0,\\ \overline{E}_k=\bigg(\frac{1}{k}\sum_{i=0}^{k-1}\varphi_i\varphi_i^{\mathrm{T}}\bigg)^{-1}, & k\ge\overline{k}_0.\end{cases}$$

For $k\ge\overline{k}_0$, notice that $E_k-\overline{E}_k=\overline{E}_k(\overline{E}_k^{-1}-E_k^{-1})E_k$ and $\lim_{k\to\infty}\big(\overline{E}_k^{-1}-E_k^{-1}\big)=\mathbf{0}$. Thus, it follows from $\sup_i|\varphi_{i,h}|\le M^0$ a.s. and Lemma 3.2 that $\lim_{k\to\infty}\big(E_k-\overline{E}_k\big)=\mathbf{0}$ for $k\ge\overline{k}_0$. This together with $\sup_i|\varphi_{i,h}|\le M^0$ a.s. implies

$$\begin{aligned}&\lim_{k\to\infty}\Bigg\{\Bigg[\bigg(\frac{1}{k}\sum_{i=0}^{k-1}\varphi_i\varphi_i^{\mathrm{T}}+\frac{1}{k\beta_0}I_{\frac{(r+r_1)(r+r_1+1)}{2}+1}\bigg)^{-1}-\bigg(\frac{1}{k}\sum_{i=0}^{k-1}\varphi_i\varphi_i^{\mathrm{T}}\bigg)^{-1}\Bigg]\\&\quad\times\frac{1}{k}\sum_{i=0}^{k-1}\varphi_i x_{i+1}^{\mathrm{T}}Px_{i+1}\Bigg\}\\&=\vec{0},\quad k\ge\overline{k}_0.\end{aligned}\tag{68}$$

It follows from $\lim_{k\to\infty}\frac{1}{k}\sum_{i=0}^{k-1}\varepsilon_i=0$ a.s., $\sup_i|\varphi_{i,h}|\le M^0$ a.s., and Lemma 3.2 that

$$\lim_{k\to\infty}\Bigg\{\bigg(\frac{1}{k}\sum_{i=0}^{k-1}\varphi_i\varphi_i^{\mathrm{T}}\bigg)^{-1}\frac{1}{k}\sum_{i=0}^{k-1}\varphi_i\varepsilon_i\Bigg\}=\vec{0},\quad k\ge\overline{k}_0.\tag{69}$$

By (67), (68), and (69), we have $\lim_{k\to\infty}\widehat{\theta}(A,B,C,V,P,k)-\theta(A,B,C,V,P)=\vec{0}$. The proof is completed. ∎

Appendix C: Proof of Theorem 4.1

Let $A=(a_{ij})\in\mathbb{R}^{r\times r}$, $P=(p_{ij})\in\mathbb{R}^{r\times r}$, and $B=(b_{ij_1})\in\mathbb{R}^{r\times r_1}$, where $i,j=1,2,\cdots,r$ and $j_1=1,2,\cdots,r_1$. By (A4) there are constants $i_0,j_0,s_0\in\{1,2,\cdots,r\}$ and $l_0\in\{1,2,\cdots,r_1\}$ such that $a_{i_0j_0}>0$ and $b_{s_0l_0}>0$.

For (29), define $\overline{G}_0:=\{\omega|\ \lim_{k\to\infty}M_k(\omega)=M\}$, and $P(\overline{G}_0)=1$. We denote $M_k(\omega):=[\mathrm{vecv}(a_1)_k(\omega)\ \ f(a_1,a_2)_k(\omega)\cdots f(a_1,b_{r_1})_k(\omega)\ \ \mathrm{vecv}(a_2)_k(\omega)\cdots\mathrm{vecv}(b_{r_1})_k(\omega)\ \ g(CVC^{\mathrm{T}})_k(\omega)]^{\mathrm{T}}$,

where

$$\begin{aligned}
&\mathrm{vecv}(a_{\overline{b}})_k(\omega) = [(a^2_{1\overline{b}})_k(\omega) \ \ (a_{1\overline{b}}a_{2\overline{b}})_k(\omega) \ \ \cdots \ \ (a^2_{r\overline{b}})_k(\omega)]^{\mathrm{T}} \in \mathbb{R}^{\frac{r(r+1)}{2}}, \quad \overline{b} = 1, 2, \cdots, r;\\
&\mathrm{vecv}(b_{\check{j}})_k(\omega) = [(b^2_{1\check{j}})_k(\omega) \ \ (b_{1\check{j}}b_{2\check{j}})_k(\omega) \ \ \cdots \ \ (b^2_{r\check{j}})_k(\omega)]^{\mathrm{T}} \in \mathbb{R}^{\frac{r(r+1)}{2}}, \quad \check{j} = 1, 2, \cdots, r_1;\\
&f(a_{\overline{b}}, b_{\check{j}})_k(\omega) := [(2a_{1\overline{b}}b_{1\check{j}})_k(\omega) \ \ (a_{1\overline{b}}b_{2\check{j}} + a_{2\overline{b}}b_{1\check{j}})_k(\omega) \ \ \cdots \ \ (2a_{r\overline{b}}b_{r\check{j}})_k(\omega)]^{\mathrm{T}} \in \mathbb{R}^{\frac{r(r+1)}{2}}
\end{aligned}$$

for $\overline{b} = 1, 2, \cdots, r,\ \check{j} = 1, 2, \cdots, r_1$;

$$f(a_{\overline{b}}, a_{\widehat{b}})_k(\omega) := [(2a_{1\overline{b}}a_{1\widehat{b}})_k(\omega) \ \ (a_{1\overline{b}}a_{2\widehat{b}} + a_{2\overline{b}}a_{1\widehat{b}})_k(\omega) \ \ \cdots \ \ (2a_{r\overline{b}}a_{r\widehat{b}})_k(\omega)]^{\mathrm{T}} \in \mathbb{R}^{\frac{r(r+1)}{2}}$$

for $\overline{b} \neq \widehat{b},\ \overline{b},\ \widehat{b} = 1, 2, \cdots, r$;

$$f(b_{\check{j}}, b_{\widehat{j}})_k(\omega) := [(2b_{1\check{j}}b_{1\widehat{j}})_k(\omega) \ \ (b_{1\check{j}}b_{2\widehat{j}} + b_{2\check{j}}b_{1\widehat{j}})_k(\omega) \ \ \cdots \ \ (2b_{r\check{j}}b_{r\widehat{j}})_k(\omega)]^{\mathrm{T}} \in \mathbb{R}^{\frac{r(r+1)}{2}}$$

for $\check{j} \neq \widehat{j},\ \check{j},\ \widehat{j} = 1, 2, \cdots, r_1$; and $g(CVC^{\mathrm{T}})_k(\omega) := [(h_{11})_k(\omega) \ \ (h_{12})_k(\omega) \ \ \cdots \ (h_{1r})_k(\omega) \ (h_{22})_k(\omega) \ \ (h_{23})_k(\omega) \ \ \cdots \ \ (h_{rr})_k(\omega)]^{\mathrm{T}} \in \mathbb{R}^{\frac{r(r+1)}{2}},\ \omega \in \overline{G}_0$. It follows from (29) and definitions of $M_k(\omega)$ and $\overline{G}_0$ that

$$\lim_{k\to\infty} \mathrm{vecv}(a_{j_0})_k(\omega) = \mathrm{vecv}(a_{j_0}), \quad \lim_{k\to\infty} \mathrm{vecv}(b_{l_0})_k(\omega) = \mathrm{vecv}(b_{l_0}), \tag{70}$$

for all $\omega \in \overline{G}_0$.

Let $\omega \in \overline{G}_0$ in the following.

By $a_{i_0j_0} > 0$, $b_{s_0l_0} > 0$, and (70), there are $k_1 < \infty$ and $k_2 < \infty$, such that $(a^2_{i_0j_0})_k(\omega) > 0$ for all $k > k_1$ and $(b^2_{s_0l_0})_k(\omega) > 0$ for all $k > k_2$, respectively.

For $j = 1, 2, \cdots, r$, the way to construct $(a_{jj_0})_k(\omega)$ and $(b_{jl_0})_k(\omega)$ is shown in Figure 2.

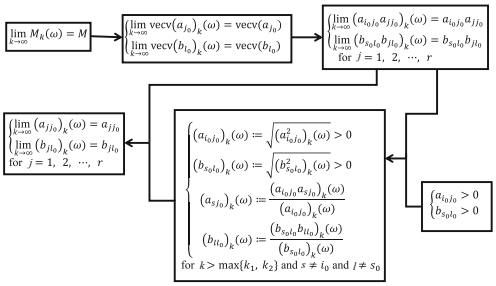


Figure 2 Design of $(a_{jj_0})_k(\omega)$ and $(b_{jl_0})_k(\omega)$

Thus, we have

$$\lim_{k\to\infty} (a_{jj_0})_k(\omega) = a_{jj_0}, \quad \lim_{k\to\infty} (b_{jl_0})_k(\omega) = b_{jl_0}, \tag{71}$$

for $j = 1, 2, \cdots, r$.

Similarly, we get

$$\begin{cases} \lim\limits_{k\to\infty} (a_{i\check{i}})_k(\omega) = a_{i\check{i}}, \quad i = 1, 2, \cdots, r; \quad \check{i} = 1, 2, \cdots, j_0 - 1, j_0 + 1, \cdots, r; \\ \lim\limits_{k\to\infty} (b_{ij^1})_k(\omega) = b_{ij^1}, \quad i = 1, 2, \cdots, r; \quad j^1 = 1, 2, \cdots, l_0 - 1, l_0 + 1, \cdots, r_1. \end{cases} \tag{72}$$

Let $A_k(\omega) := \big((a_{ij})_k(\omega)\big) \in \mathbb{R}^{r\times r},\ B_k(\omega) := \big((b_{ij_1})_k(\omega)\big) \in \mathbb{R}^{r\times r_1}$. It follows from (71), (72), and definitions of $A_k(\omega)$ and $B_k(\omega)$ that $\lim_{k\to\infty} A_k(\omega) = A,\ \lim_{k\to\infty} B_k(\omega) = B$.

Now, (30) in Algorithm 1 is equivalent to

$$\begin{aligned}\widehat{P}_{k+1}(\omega) = A_{k+1}^{\mathrm{T}}(\omega)\widehat{P}_k(\omega)A_{k+1}(\omega) - A_{k+1}^{\mathrm{T}}(\omega)\widehat{P}_k(\omega)B_{k+1}(\omega)\big(R + B_{k+1}^{\mathrm{T}}(\omega)\widehat{P}_k(\omega)\\ \times B_{k+1}(\omega)\big)^{-1}B_{k+1}^{\mathrm{T}}(\omega)\widehat{P}_k(\omega)A_{k+1}(\omega) + Q,\end{aligned} \tag{73}$$

where $k+1 > \overline{k}_1 = \max\{k_1, k_2\}$, $\widehat{P}_{\overline{k}_1}(\omega) > 0$, and $\overline{k}_1 < \infty$.

By (73) and mathematical induction, we know that $\widehat{P}_k(\omega) > 0$ for $k \geq \overline{k}_1$.

Now, Algorithm 1 is summarized in (73), and $\widehat{P}_{\overline{k}_1}(\omega) > 0$. Thus, by Theorem 3.4 of [36], we get $\lim_{k\to\infty}\widehat{P}_k(\omega) = P^*$ and $\lim_{k\to\infty}\widehat{K}_k(\omega) = K^*$, where $\omega \in \overline{G}_0$. The proof is completed. ∎

Appendix D: Proof of Theorem 4.2

By the proof idea of Theorem 8.3 in [36], we can proof Theorem 4.2.

For (31), we have $\widehat{K}_k \in \mathcal{F}_k$, which together with (32) implies $\widehat{u}_k^0 \in \mathcal{F}_k$. We know that $\lim_{k\to\infty}\widehat{K}_k(\omega) = K^*$ by Theorem 4.1, where $\omega \in \overline{G}_0$ and $P(\overline{G}_0) = 1$.

Now, we show that $\{\widehat{u}_k^0\} \in \mathcal{U}_{ad}$. Since $\lim_{k\to\infty}\widehat{K}_k(\omega) = K^*$ and $A - BK^*$ is the stable matrix. By Theorem 8.3 of [36], we get

$$\|(A - B\widehat{K}_k(\omega))(A - B\widehat{K}_{k-1}(\omega))\cdots(A - B\widehat{K}_0(\omega))\| < c\varrho^{k+1}, \varrho \in (0,1), \quad \forall k \geq 0, \tag{74}$$

where c is a positive constant and $\omega \in \overline{G}_0$ and $P(\overline{G}_0) = 1$.

It follows from (1) and (32) that

$$\begin{aligned}x_k^0 = (A - B\widehat{K}_{k-1})(A - B\widehat{K}_{k-2})\cdots(A - B\widehat{K}_0)x_0^0 + \sum_{i=1}^{k-1}\big[(A - B\widehat{K}_{k-1})\\ \times(A - B\widehat{K}_{k-2})\cdots(A - B\widehat{K}_{k-i})C\omega_{k-i}\big] + C\omega_k.\end{aligned} \tag{75}$$

It follows from (74), (75), and a.s. boundedness of $\{\omega_k\}_{k=0}^{\infty}$ that there exists a constant $M^1 > 0$ such that $\sup_k \|x_k^0\| \leq M^1$ a.s.. This together with (32) implies $\{\widehat{u}_k^0\} \in \mathcal{U}_{ad}$.

Now, we prove that $\widehat{u}_k^0$ is optimal. By Theorem 8.3 of [36], for any $\breve{u} \in \mathcal{U}_{ad}$, we have

$$\begin{aligned}J(\breve{u}) = \mathrm{tr}(P^*CVC^{\mathrm{T}}) + \limsup_{k\to\infty}\Bigg(\frac{1}{k}\sum_{i=0}^{k-1}\big[\breve{u}_i + (R + B^{\mathrm{T}}P^*B)^{-1}B^{\mathrm{T}}P^*Ax_i\big]^{\mathrm{T}}\\ \times(R + B^{\mathrm{T}}P^*B)\big[\breve{u}_i + (R + B^{\mathrm{T}}P^*B)^{-1}B^{\mathrm{T}}P^*Ax_i\big]\Bigg).\end{aligned} \tag{76}$$

Thus to complete the proof it suffices to show that for the control law $\widehat{u}_k^0$ defined by (32)

$$\limsup_{k\to\infty}\frac{1}{k}\sum_{i=0}^{k-1}\|\widehat{u}_i^0 + (R + B^{\mathrm{T}}P^*B)^{-1}B^{\mathrm{T}}P^*Ax_i^0\|^2 = 0. \tag{77}$$

It follows from $\lim_{k\to\infty}\widehat{K}_k = K^*$ a.s. and $\sup_k\|x_k^0\| \leq M^1$ a.s. that (77) is true. Hence, $\widehat{u}_k^0$ is the optimal controller. The proof is completed. ∎