International Journal of Control, Automation, and Systems 21(9) (2023) 3068-3079
http://dx.doi.org/10.1007/s12555-022-0355-6
ISSN:1598-6446 eISSN:2005-4092
http://www.springer.com/12555

Reinforcement Learning for Input Constrained Sub-optimal Tracking Control in Discrete-time Two-time-scale Systems

Xuejie Que, Zhenlei Wang*, and Xin Wang

Abstract: Two-time-scale (TTS) systems were proposed to describe accurately complex systems that include multiple variables running on two-time scales. Different response speeds of variables and incomplete model information affect the tracking performance of TTS systems. For tracking control of an unknown model, the practicability of reinforcement learning (RL) has been subject to criticism, as the method requires a stable initial policy. Based on singular perturbation theory (SPT), a composite sub-optimal tracking policy is investigated combining model information with measured data. Besides, a selection criterion for the initial stabilizing policy is presented by considering the policy as an input constraint. The proposed method integrating RL technique with convex optimization improves the tracking performance and practicability effectively. Finally, an emulation experiment in F-8 aircraft is given to demonstrate the validity of the developed method.

Keywords: Convex optimization, input constrained, reinforcement learning, sub-optimal tracking control, two-time-scale system.

1. INTRODUCTION

Two-time-scale characteristics of complex systems were described accurately by TTS systems, such as aircraft, network control and process industry [1-3]. Take the aircraft for example, the characteristics is embodied in a slow phugoid mode and a fast short-period mode, where change in position of mass center is slow and change in angle of attack is fast [1]. The TTS systems not only have characteristics of multi-variable, but also have different response speeds between slow variables and fast variables. Curse of dimensionality and ill-condition may generate unacceptable computational complexity and performance deficiency in performing controller design, respectively [4,5]. That is, fast variables cannot respond immediately to controller designing in slow time scale, and slow variables have a little response to controller designing in fast time scale. The SPT was presented to counteract the problems as the TTS systems can be converted to singularly perturbed systems in form [6]. Nonetheless, most of the existing achievements rely so heavily on model information [7,8]. Model-free method requires a large amount of data and has poor ability to explain the influence of variables from inside [9,10]. In practice, models of the TTS systems can be established but incomplete due to factors of unknown mechanisms and technology. So, it is worth exploring that the design of controller combining model information with measured data, which ensures reliable performance as far as possible while providing internal interpretation.

RL as a machine learning method has been effectively employed to find the optimal control policy under the unknown model information [11]. A variety of RL techniques, such as Q-learning, actor-critic and adaptive dynamic programming, were widely applied in control field [12-16]. Recently, a composite sub-optimal control strategy was developed for a class of continuous-time TTS systems [17]. In addition, reduced-dimensional RL technique was applied to explore optimal problem with two-time-scale [18]. There is a common requirement in aforementioned techniques: stable policy was regarded as an initial condition. However, it is difficult to select the initial stabilizing policy when the model information is incomplete.

On the other hand, the optimal tracking control (OTC) problem for TTS system continues to be ongoing arguments. It is aimed at designing a tracking controller such that a reference trajectory is tracked by output in an optimal manner [19-22]. Authors of [23,24] discussed the OTC for flotation industrial process on two-time-scales and presented a dual-rate data-driven algorithm by mean

Manuscript received April 28, 2022; revised September 3, 2022; accepted October 11, 2022. Recommended by Associate Editor Jun Moon under the direction of Senior Editor Jong Min Lee. This work was supported by National Natural Science Foundation of China (Basic Science Center Program: 61988101), Natural Science Foundation of China (62233005, 62273149), Fundamental Research Funds for the Central Universities and Shanghai AI Lab.

Xuejie Que and Zhenlei Wang are with the Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China (e-mails: {xuejie_que, wangzhen_l}@163.com). Xin Wang is with the Center of Electrical and Electronic Technology, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: wangxin26@sjtu.edu.cn).
* Corresponding author.

of lifting technology. However, when the desired trajectory is not static, the OTC problem has not been addressed adequately. In addition, there is a challenge to analyze the asymptotic behavior of the TTS system under single rate sampling. Taking discrete-time TTS model under fast sampling as an example, it is no problem to analyze the asymptotic behavior of the fast subsystem, but accurately asymptotic stability of the slow subsystem can not be characterized in the fast time scale. To avoid misjudgment of stability, the slow subsystem in the slow time scale is studied to characterize its asymptotic stability.

This paper is motivated the fact that the selection of the initial stabilizing policy has difficulties and tracking performance is affected by time scale in TTS systems. It aims to develop a composite algorithm based on RL technique and convex optimization to learn the sub-optimal tracking control solution for TTS systems. First, fast subsystem and slow subsystem are derived from a high-order TTS system using SPT. To improve tracking performance, the desired trajectory is tracked by the slow subsystem, the fast subsystem is required to reduce oscillation. Combining LMI technique with Lyapunov stability theorem (LST) [25], the stabilizing policy constraints are encoded to two subproblems. It provides a choice of initial stabilizing policy. Then, a composite sub-optimal algorithm is proposed by using of RL technique and convex optimization, which not relies on Hamiltonian. Finally, the asymptotic stability for tracking error system in the slow time scale is analyzed under [26].

Compared with the existing literatures [4,7,10], advantages of this paper are roughly summarized in the following three points:

1) Based on SPT, the proposed method combining data with existing model information guarantees the accuracy of results and reduces dependence on model. It is stressed that system performance is affected by time scale. Moreover, lifting technique and inductive reasoning are employed to research the asymptotic behavior of tracking error system in the slow time scale.
2) In TTS systems, the exploration of the initial stabilizing policy in RL method provides an idea for its selection, which improves the practicability of the RL method. Furthermore, the interpretability of learning policy is improved by taking full advantage of available model information.
3) A new RL method integrating SPT with Lagrangian duality theory is developed. Unlike traditional RL approach, policy evaluation and policy improvement depend on Karush-Kuhn-Tucker (KKT) conditions rather than Hamiltonian.

This article is organized as follows: Section 2 uses SPT to separate the TTS system. In Section 3, the input constrained OTC problem and two constrained optimization sub-problems are formulated. Section 4 gives a composite sub-optimal tacking controller combined Q-learning with policy iteration to track the desired trajectory. Section 5 analyzes system asymptotic stability, as well as algorithm convergence. In Section 6, an emulation experiment in F-8C aircraft is considered to demonstrate the validity of the developed method. Conclusions are presented in Section 7.

Notation: Symbol T that appears throughout the article stands for matrix transposition; matrix inequality $P > 0$ and $P \geq 0$ indicate positive definite matrices, positive semi-definite matrices, respectively; the element $*$ under the main diagonal of the symmetric matrix denotes an ellipsis for terms that are induced by symmetry; Tr represents trace of matrix ; ρ stands for spectral radius. $\otimes$ denotes the Kronecker product; $vec(H)$ is a column vector formed by stacking the columns of matrix H.

2. TTS SYSTEM DESCRIPTION AND DECOMPOSITION

In this section, a class of linear discrete-time TTS systems with partially unknown dynamics are discussed. Based on SPT, the high-order TTS system is separated into the corresponding fast dynamics and the slow dynamics.

2.1. TTS system description

Consider the fast sampling linear discrete-time TTS system

$$\begin{aligned} x_1(k+1) &= (I+\varepsilon A_1)x_1(k)+\varepsilon A_2 x_2(k)+\varepsilon B_1 u(k),\\ x_2(k+1) &= A_3 x_1(k)+A_4 x_2(k)+B_2 u(k),\\ y(k) &= C_1 x_1(k)+C_2 x_2(k), \end{aligned} \tag{1}$$

where $x_1 \in \mathbb{R}^{m_1}$ is the slow state vector, and $x_2 \in \mathbb{R}^{m_2}$ is the fast state vector; σ and $\tilde{\sigma}$ are the initial state of x_1 and x_2 respectively. Different response speeds of the TTS systems are caused by singularly perturbation parameter ε that affects positions of eigenvalues in the unit circle, where positive scalar ε is far less than 1. $u \in \mathbb{R}^q$ is the control input vector; $y \in \mathbb{R}^p$ is the system output; k represents the fast time scale; the knowledge of the system is partially unknown, where A_3, A_4, B_2, C_1 and C_2 are known matrices with appropriate dimensions, and A_1, A_2, B_1, ε are unknown.

Assumption 1: The matrix $I-A_4$ is nonsingular.

Note that Assumption 1 is essential for separating the discrete-time TTS system (1). Because the standard requirement of the system relies on $I-A_4$, that is, the inexistence of isolated root if the matrix is singular. The existence of isolated roots ensures that slow subsystem is well defined.

2.2. Time-scale decomposition

Slow subsystem and fast subsystem are defined based on Assumption 1 and SPT. The slow subsystem is equiv-

alent to quasi-steady state model, which is obtained by replacing fast state vector with its steady-state algebraic equation. The fast subsystem, also called boundary layer model, is deviation between the quasi-steady state model and the whole order model.

Neglecting the fast mode in the whole order system (1), the slow mode can be described as follows:

$$\begin{aligned} x_{1s}(k+1) &= (I+\varepsilon A_1)x_{1s}(k)+\varepsilon A_2 x_{2s}(k)+\varepsilon B_1 u_s(k),\\ x_{2s}(k+1) &= A_3 x_{1s}(k)+A_4 x_{2s}(k)+B_2 u_s(k),\\ y(k) &= C_1 x_1(k)+C_2 x_2(k), \end{aligned} \tag{2}$$

note $A_{1\varepsilon}=(I+\varepsilon A_1)$. We consider slow mode x_{2s} in the fast time scale k as a constant, which is the slow part corresponding fast state vector x_2. The steady-state algebraic equation with respect to the second equation of (2) is expressed as

$$x_{2s}(k)=(I-A_4)^{-1}\left[A_3 x_{1s}(k)+B_2 u_s(k)\right]. \tag{3}$$

There is no fast part in slow state vector x_1 because fast mode is neglected. Substituting (3) into (2), the quasi-steady state model is obtained

$$\begin{aligned} x_s(k+1) &= \mathcal{A}_s(k)x_s(k)+\mathcal{B}_s u_s(k),\\ y_s(k) &= \mathcal{C}_s x_s(k)+\mathcal{D}_s u_s(k), \end{aligned} \tag{4}$$

where $\mathcal{A}_s = I+\varepsilon[A_1+A_2(I-A_4)^{-1}A_3]$, $\mathcal{B}_s=\varepsilon[B_1+A_2(I-A_4)^{-1}B_2]$, $\mathcal{C}_s=C_1+C_2(I-A_4)^{-1}A_3$, $\mathcal{D}_s=C_2(I-A_4)^{-1}B_2$. Combining the quasi-steady state model (4) with the whole order system (1), the fast subsystem is obtained as follows:

$$\begin{aligned} x_f(k+1) &= \mathcal{A}_f x_{2f}(k)+\mathcal{B}_f u_f(k),\\ y_f(k) &= \mathcal{C}_f x_f(k), \end{aligned} \tag{5}$$

where $\mathcal{A}_f=A_4$, $\mathcal{B}_f=B_2$, $\mathcal{C}_f=C_2$, $x_{2f}=x_2-x_{2s}$.

For example, the tracking performance of aircraft is mainly affected by incremental pitch attitude κ and incremental velocity υ, and level of oscillation is related to pitch rate q and incremental angle of attack α. Corresponding to the system (1), $x_1=\varpi_1-N\varpi_2$, $x_2=\varpi_2=\begin{bmatrix}\alpha^T & q^T\end{bmatrix}^T$, $\varpi_1=\begin{bmatrix}\upsilon^T & \kappa^T\end{bmatrix}^T$, $u=\delta_e$ is elevator position, N is a matrix with appropriate dimensions. Ill-condition is caused by different response speeds between ϖ_1 and ϖ_2. To solve the problem, the slow subsystem $x_{1s}=\varpi_{1s}+\bar{N}\varpi_{1s}+\tilde{N}u_s$ is obtained by replacing ϖ_{2s} with its steady-state algebraic equation, and the fast subsystem $x_f=\varpi_2-\varpi_{2s}$ is presented. The computational difficulty of tracking problem is decreased by reducing model order and the ill-condition can be solved by designing controllers in different time-scale.

3. PROBLEM FORMULATION

In this section, a linear command generator system is presented to generate reference trajectory. Then, we formulate the stabilizing policies as input constraints. As the original system (1) decomposes, the input constrained OTC problem and input constrained the optimal control (OC) problem are proposed. It provides ideas for initial stabilizing and tracking performance improvement.

3.1. Linear command generator

It is assumed that the reference trajectory dynamics is described by a linear command generator

$$r(k+1)=Lr(k), \tag{6}$$

where $r\in\mathbb{R}^p$ is the dynamical trajectory vector; δ is the initial state $r(0)$; L is an unknown square matrix with appropriate dimensions and $\rho(L)<1$.

3.2. Constrained optimal tracking control problem

The performance index J_u is denoted as follows:

$$\begin{aligned} J_u = \sum_{k=i}^{\infty}\gamma^{k-i}&(y_s(k)-r(k))^T Q(y_s(k)-r(k))\\ &+(y_f(k))^T Q y_f(k)+u(k)^T R u(k), \end{aligned} \tag{7}$$

where $Q\geq 0, R>0, y(0)=\tau$. $0<\gamma\leq 1$ is discount factor. We formulate the input constrained OTC problem as

Primal problem I (OTC problem for the original system)**:**

$$\hat{J}_u=\min_{u\in\mathfrak{U}} J_u, \tag{8}$$

$$\begin{aligned} \text{s.t. } & x_s(k+1)=\mathcal{A}_s(k)x_s(k)+\mathcal{B}_s u_s(k),\\ & x_f(k+1)=\mathcal{A}_f x_{2f}(k)+\mathcal{B}_f u_f(k),\\ & y_s(k)=\mathcal{C}_s x_s(k)+\mathcal{D}_s u_s(k),\\ & y_f(k)=\mathcal{C}_f x_f(k),\\ & r(k+1)=Lr(k), \end{aligned} \tag{9}$$

where $\mathfrak{U}$ is a set of stabilizing policy.

In problem formulation, Primal problem I is similar with the standard linear model predictive control (MPC) [27,28]. However, the difference can be seen at input constraint and performance index. An implicit constraint $u\in\mathfrak{U}$ is proposed in Primal problem I, instead of explicit constraint $u_{\min}\leqslant u_{k+j}\leqslant u_{\max}$. Discount factor γ is added to avoid that the value of the performance index goes to infinity. Furthermore, the linear MPC controller relies so heavily on model information compared with Primal problem I.

3.3. Constrained optimization sub-problems

On the basis of the time-scale separation property, the OTC achieved by the slow dynamics and the oscillation is restrained effectively by the fast dynamics. The aim of the constrained OTC problem is to find the optimal tracking policy u^* from feasible set so as to the output $y_s(k)$ tracks the desired trajectory $r(k)$. The structure of the tracking control problem is displayed in Fig. 1.

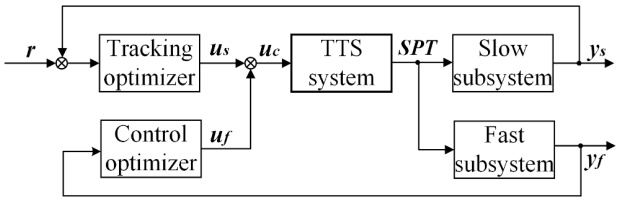

Fig. 1. Block diagram of OTC.

An augmented system is constructed by the slow dynamics (4) and the desired trajectory system (6) before formulating Sub-problem 1a

$$\begin{aligned}\xi(k+1) &= \mathcal{A}_L\xi(k) + \widehat{\mathcal{B}}u_s(k),\\ y_s(k) &= \widehat{\mathcal{C}}\xi(k) + \mathcal{D}_s u_s(k),\end{aligned} \tag{10}$$

where $\xi(k) = \begin{bmatrix} x_s(k) \\ r(k) \end{bmatrix}$, $\mathcal{A}_L = \begin{bmatrix} \mathcal{A}_s & 0 \\ 0 & L \end{bmatrix}$, $\widehat{\mathcal{B}} = \begin{bmatrix} \mathcal{B}_s \\ 0 \end{bmatrix}$, $\widehat{\mathcal{C}} = \begin{bmatrix} \mathcal{C}_s^T \\ 0 \end{bmatrix}^T$, $\xi(0) = \begin{bmatrix} \sigma \\ \delta \end{bmatrix} = \eta$.

Designing a state feedback-based policy $u_s(k)$ for the augmented system (10)

$$u_s(k) = -F_s\xi(k) = -F_1x_s(k) - F_2r(k). \tag{11}$$

Another augmented system is composed of the augmented system (10) and tracking policy (11)

$$\begin{bmatrix} \xi(k+1) \\ u_s(k+1) \end{bmatrix} = \begin{bmatrix} \mathcal{A}_L & \widehat{\mathcal{B}} \\ -F_s\mathcal{A}_L & -F_s\widehat{\mathcal{B}} \end{bmatrix} \begin{bmatrix} \xi(k) \\ u_s(k) \end{bmatrix}. \tag{12}$$

Primal sub-problem 1a (Input constrained OTC problem)**:** $\mathcal{F}_s$ is a set which contains stabilizing state feedback gains for the slow subsystem in slow time scale n, where the expression of $\mathcal{F}_s$ is presented in the form of (62). The input constrained OTC sub-problem is described as

$$\hat{J}_{F_s} = \min_{F_s} J_{F_s}, \tag{13}$$

$$\text{s.t. } \xi(k+1) = \mathcal{A}_L\xi(k) + \widehat{\mathcal{B}}u_s(k), \tag{14}$$

$$F_s \in \mathcal{F}_s, \tag{15}$$

where $J_{F_s} = \sum_{k=i}^{\infty} \gamma^{k-i} \begin{bmatrix} \xi(k) \\ -F_s\xi(k) \end{bmatrix}^T \Phi_s \begin{bmatrix} \xi(k) \\ -F_s\xi(k) \end{bmatrix}$, $\Phi_s = \begin{bmatrix} \hat{Q}_s & \tilde{Q}_s \\ * & \hat{R} \end{bmatrix}$, $\hat{Q}_s = \mathcal{C}^T Q \mathcal{C}$, $\tilde{Q}_s = \mathcal{C}^T Q \mathcal{D}_s$, $\hat{R} = \mathcal{D}_s^T Q \mathcal{D}_s + R$, $\mathcal{C} = \begin{bmatrix} \mathcal{C}_s & -I \end{bmatrix}$. Let F_s^* be the optimal feedback gain for the input constrained OTC, it can be defined as $F_s^* = arg\min_{F_s \in \mathcal{F}_s} J_{F_s}$.

Lemma 1: Let $\mathcal{A}_F = \mathcal{A}_s - \mathcal{B}_s F_1$, there exists a nonsingular matrix $\tilde{\Psi} = \begin{bmatrix} I & 0 \\ -F_s & I \end{bmatrix}$, such that $\rho(\Gamma_{F_S}) < 1$ is equivalent to $\rho(A_F) < 1$.

Primal sub-problem 1b (Input constrained OC problem)**:** Choosing a state feedback control policy $u_f(k) = -F_f x_f(k)$, the optimal control problem can be formulated as

$$\hat{J}_{F_f} = \min_{F_f} J_{F_f}, \tag{16}$$

$$\text{s.t. } x_f(k+1) = \mathcal{A}_f x_{2f}(k) + \mathcal{B}_f u_f(k),$$

$$F_f \in \mathcal{F}_f, \tag{17}$$

where $J_{F_f} = \sum_{k=i}^{\infty} \gamma^{k-i} \begin{bmatrix} x_f(k) \\ -F_f x_f(k) \end{bmatrix}^T \Phi_f \begin{bmatrix} x_f(k) \\ -F_f x_f(k) \end{bmatrix}$, $\Phi_f = \begin{bmatrix} Q & 0 \\ * & R \end{bmatrix}$, $\hat{\varsigma}_f(k) = \begin{bmatrix} x_f(k) \\ -F_f x_f(k) \end{bmatrix}$. Similarly, define $F_f^* = arg\min_{F_f \in \mathcal{F}_f} J_{F_f}$ as the optimal feedback gain for the input constrained OC.

In order to provide viable selection criteria of the initial stabilizing policies, the stabilizing policies are limited within two sets of stabilizing feedback gains. The challenges include two aspects: the form of constraint not conform to standardized optimization problem; the elements of the sets are hard to get. The explanation is presented the following section.

4. SUB-OPTIMAL TRACKING CONTROLLER

In this section, standardized forms of sub-problems 1a and 1b are provided to solve the problem of the initial stabilizing policy. Then, dual sub-problems are obtained by using convex optimization. Strong duality between the dual problems and the primal sub-problems is presented to guarantee the equivalence of the global optimal solutions. Based on RL technique, a composite sub-optimal algorithm is obtained. The overall flow is summarized in Fig. 2.

Lemma 2 [29]**:** Suppose that $S \geq 0$ is a partitioned

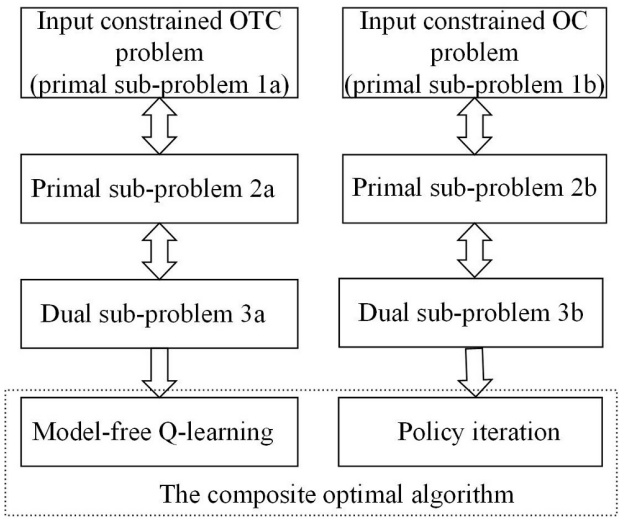

Fig. 2. Block diagram of sub-optimal tracking controller.

square matrix as follows: $S=\begin{pmatrix} S_{11} & S_{12} \\ S_{12}{}^T & S_{22} \end{pmatrix}$, where S_{11} symmetric matrices, $S_{22}>0$. Then

$$\begin{bmatrix} I \\ -F \end{bmatrix}^T S \begin{bmatrix} I \\ -F \end{bmatrix} \geqslant S_{11}-S_{12}S_{22}^{-1}S_{12}^T, \tag{18}$$

where $S_{11}-S_{12}S_{22}^{-1}S_{12}^T=\begin{bmatrix} I \\ -S_{22}^{-1}S_{12}^T \end{bmatrix}^T S \begin{bmatrix} I \\ -S_{22}^{-1}S_{12}^T \end{bmatrix}$. The equality in (18) holds if and only if $F=S_{22}^{-1}S_{12}^T$.

4.1. Inequality constrained optimization sub-problems

Two Primal sub-problems 2a and 2b, which are equivalent to Primal sub-problems 1a and 1b, are presented by means of spectral decomposition method and LMI technique. The purpose of the operation is to give standard forms of convex optimization problems. Meanwhile, the selection criteria of the initial stabilizing policies are obtained, which depend on positive semi-definite matrices composed of data. Moreover, we give two Lagrangian functions referring to Primal sub-problems 2a and 2b.

Primal sub-problem 2a (Inequality constrained OTC problem)**:**

$$J_\Theta=\min_\Theta tr(\Phi_s\Theta), \tag{19}$$

$$\text{s.t. } \gamma\Gamma_{F_s}\Theta\Gamma_{F_s}{}^T+\begin{bmatrix} I_{m_1+p} \\ -F_s \end{bmatrix}\Pi\begin{bmatrix} I_{m_1+p} \\ -F_s \end{bmatrix}^T=\Theta, \tag{20}$$

$$\Theta\geq 0, \tag{21}$$

where $\Theta=\sum_{k=i}^{\infty}\gamma^{k-i}\begin{bmatrix} \xi(k) \\ -F_s\xi(k) \end{bmatrix}\begin{bmatrix} \xi(k) \\ -F_s\xi(k) \end{bmatrix}^T$, $\Pi=\eta\eta^T>0$.

Proof: Using spectral decomposition method, there exists an orthogonal matrix $\mathcal{Q}$ and a diagonal matrix Ξ_s such that $\Phi_s=\mathcal{Q}\Xi_s\mathcal{Q}^T$. Substituting for Ξ_s in (13), one has

$$\hat{J}_{F_s}=\min_{F_s}\sum_{k=i}^{\infty}\gamma^{k-i}\begin{bmatrix} \xi(k) \\ -F_s\xi(k) \end{bmatrix}^T\mathcal{Q}\Xi_s\mathcal{Q}^T\begin{bmatrix} \xi(k) \\ -F_s\xi(k) \end{bmatrix}. \tag{22}$$

Putting (14) in (22) and applying LMI technique yields

$$\hat{J}_{F_s}=\min_{F_s}Tr\sum_{k=i}^{\infty}\{\gamma^{k-i}\Xi_s\mathcal{Q}^T \times\begin{bmatrix} \xi(k) \\ -F_s\xi(k) \end{bmatrix}\begin{bmatrix} \xi(k) \\ -F_s\xi(k) \end{bmatrix}^T\mathcal{Q}\Big\}. \tag{23}$$

Note $\Theta=\sum_{k=i}^{\infty}\gamma^{k-i}\begin{bmatrix} \xi(k) \\ -F_s\xi(k) \end{bmatrix}\begin{bmatrix} \xi(k) \\ -F_s\xi(k) \end{bmatrix}^T$, one can write (23) as

$$\hat{J}_{F_s}=\min_\Theta Tr(\Phi_s\Theta), \tag{24}$$

where

$$\begin{aligned}\Theta&=\sum_{k=i}^{\infty}\gamma^{k-i}\begin{bmatrix} \xi(k) \\ -F_s\xi(k) \end{bmatrix}\begin{bmatrix} \xi(k) \\ -F_s\xi(k) \end{bmatrix}^T\\&=\gamma\Gamma_{F_s}\Theta\Gamma_{F_s}{}^T+\begin{bmatrix} I \\ -F_s \end{bmatrix}\Pi\begin{bmatrix} I \\ -F_s \end{bmatrix}^T.\end{aligned}$$

Primal sub-problem 1a is identical to Primal sub-problem 2a. □

An implicit constraint $F_s\in\mathcal{F}_s$ is equivalent to an inequation $\Theta\geq 0$ based on LST and Schur Complement [30]. According to the formulation of Θ, the selection criterion of initial stabilizing policy under unknown model information are converted into $\Theta\geq 0$.

Then, Lagrangian function of Primal sub-problem 2a can be formulated as

$$\begin{aligned}L_s(P_s,P_0,F_s,\Theta)=&Tr(\Phi_s\Theta)+Tr\left[\left(\gamma\Gamma_{F_s}\Theta\Gamma_{F_s}^T-\Theta\right)P_s\right]\\&+Tr\left(\begin{bmatrix} I_{m_1+p} \\ -F_s \end{bmatrix}\Pi\begin{bmatrix} I_{m_1+p} \\ -F_s \end{bmatrix}^T P_s\right)\\&-Tr\left(P_0\Theta\right),\end{aligned} \tag{25}$$

where Lagrange multiplier P_s is associated with the equality constraint (20) and Lagrange multiplier P_0 is associated with inequality constraint (21). $L_s(P_s, P_0, F_s, \Theta)\geq 0$.

Primal sub-problem 2b (Inequality constrained OC problem)**:**

$$J_\Psi=\min_\Psi tr(\Phi_f\Psi), \tag{26}$$

$$\text{s.t } \gamma\Gamma_{F_f}\Psi\Gamma_{F_f}{}^T+\begin{bmatrix} I_{m_2} \\ -F_f \end{bmatrix}\chi\begin{bmatrix} I_{m_2} \\ -F_f \end{bmatrix}^T=\Psi,$$

$$\Psi\geq 0, \tag{27}$$

where $\Psi=\sum_{k=i}^{\infty}\gamma^{k-i}\begin{bmatrix} x_f^k(k) \\ -F_fx_f^k(k) \end{bmatrix}\begin{bmatrix} x_f^k(k) \\ -F_fx_f^k(k) \end{bmatrix}^T$, $\chi=\upsilon\upsilon^T$. Lagrange function

$$\begin{aligned}L_f(P_f,\,P_1,F_s,\Psi)=&Tr\left\{\begin{bmatrix} I_{m_2} \\ -F_f \end{bmatrix}\chi\begin{bmatrix} I_{m_2} \\ -F_f \end{bmatrix}^T P_f\right\}\\&+Tr\left\{\left(\gamma\Gamma_{F_f}P_f\Gamma_{F_f}{}^T+\Phi_f-P_f\right)\Psi\right\}\\&-Tr(P_1\Psi).\end{aligned} \tag{28}$$

4.2. Dual sub-problems

In this subsection, Dual sub-problems 3a and 3b are proposed on the basis of convex optimization. It aims to give the global optimal policies of the sub-problems. Then, strong duality between the dual sub-problems and the primal sub-problems is proved by using LMI technique and LST.

Dual sub-problem 3a:

$$J_s^d=\sup_{P_s,P_0}d_s\left(P_s,\,P_0\right), \tag{29}$$

where Lagrange dual function $d_s(P_s,\ P_0)$ is the lower bound of Lagrangian function, $P_0 \geqslant 0$ and note

$$d_s(P_s,\ P_0) = \inf_{F_s \in \mathcal{F}_s} L_s(P_s,\ P_0,\ F_s,\ \Theta). \tag{30}$$

Dual Sub-problem 3b:

$$J_f^d = \sup_{P_f, P_1} d_f(P_f,\ P_1), \tag{31}$$

where $d_f(P_f,\ P_1) = \inf_{F_f \in \mathcal{F}_f} L_f(P_f,\ P_1,\ F_f,\ \Psi)$ is Lagrange dual function, $P_1 \geqslant 0$.

Theorem 1: There is a strong duality between Primal sub-problem 2a and Dual sub-problem 3a, that is $J_s^d = J_\Theta$.

Proof: The relationship between Lagrange dual function and Lagrangian function means that the inequality $J_s^d \leq J_\Theta$ follows. To derive $J_s^d = J_\Theta$, it remains to show that inequality $J_\Theta \leq J_s^d$ holds.

We first rewrite Primal sub-problem 1a as

$$\hat{J}_\Theta = \min_{F_s} Tr\left(\begin{bmatrix} I_{m_1+p} \\ -F_s \end{bmatrix} \Pi \begin{bmatrix} I_{m_1+p} \\ -F_s \end{bmatrix}^T P_s \right), \tag{32}$$

$$\text{s.t. } \gamma \Gamma_{F_s}{}^T P_s \Gamma_{F_s} + \Phi_s = P_s,$$
$$F_s \in \mathcal{F}_s, \tag{33}$$

where $\Gamma_{F_s} = \begin{bmatrix} \mathcal{A}_L & \widehat{\mathcal{B}} \\ -F_s\mathcal{A}_L & -F_s\widehat{\mathcal{B}} \end{bmatrix}$, $P_s = \sum_{k=i}^{\infty} \gamma^{k-i} \{\Gamma_{F_s}{}^k\}^T \mathcal{Q} \times \Xi_s \mathcal{Q}^T \Gamma_{F_s}{}^k$. The corresponding Lagrange function is

$$L_s(P_s,\ P_0, F_s) = Tr\left\{ \begin{bmatrix} I_{m_1+p} \\ -F_s \end{bmatrix} \Pi \begin{bmatrix} I_{m_1+p} \\ -F_s \end{bmatrix}^T P_s \right\} + Tr\left\{ \left(\gamma \Gamma_{F_s} P_s \Gamma_{F_s}{}^T + \Phi_s - P_s \right) P_0 \right\}. \tag{34}$$

We have the Lagrange dual function $\hat{d}_s(P_s, P_0)$ of the problem $\hat{J}_\Theta$

$$\hat{d}_s(P_s) = \inf_{F_s \in \mathcal{F}_s} \hat{L}(P_s,\ F_s) = \begin{cases} \inf_{F_s \in \mathcal{F}_s} \Delta, & \text{if } (P_s, P_0) \in \mathfrak{B}, \\ -\infty, & \text{otherwise}, \end{cases} \tag{35}$$

where $\mathfrak{B} = \left\{ P_s : \gamma \Gamma_{F_s}{}^T P_s \Gamma_{F_s} + \Phi_s - P_s \geq 0,\ F_s \in \mathcal{F}_s \right\}$, $\Delta = Tr\left\{ \begin{bmatrix} I_{m_1+p} \\ -F_s \end{bmatrix} \Pi \begin{bmatrix} I_{m_1+p} \\ -F_s \end{bmatrix}^T P_s \right\}$. Lagrange dual problem is $\hat{J}_s^d = \sup_{P_s} \hat{d}_s(P_s)$. Primal sub-problem 1a and Lagrangain function $\hat{L}_s(P_s,\ P_0,\ F_s)$ are equivalent to the problem $\hat{J}_\Theta$ and $L_s(P_s,\ P_0,\ F_s,\ \Theta)$, respectively. $J_\Theta \leqslant J_s^d$ can be achieved by $\hat{J}_\Theta \leqslant \hat{J}_s^d$.

Obviously, $\hat{d}_s(P_s) \leqslant \hat{J}_s^d$. We next show that the equality $\hat{J}_\Theta = \hat{d}_s(P_s)$ holds. Suppose that $\mathfrak{P}$ is nonempty, then $\mathcal{P} \in \mathfrak{P}$, if and only if there exists a Lagrange multiplier $\mathcal{P}$ such that the inequality holds for each $F_s \in \mathcal{F}_s$

$$\gamma \Gamma_{F_s} \mathcal{P} \Gamma_{F_s}{}^T + \Phi_s - \mathcal{P} \geq 0. \tag{36}$$

Putting Γ_{F_s} in (36), one has

$$\gamma \Gamma_{F_s}{}^T \mathcal{P} \Gamma_{F_s} + \Phi_s = \gamma \Im + \Phi_s \geqslant \mathcal{P} = \gamma \wp + \Phi_s, \tag{37}$$

where

$$\Im = \begin{bmatrix} \mathcal{A}_L & \widehat{\mathcal{B}} \end{bmatrix}^T \begin{bmatrix} I_{m_1+p} \\ -F_s \end{bmatrix}^T \mathcal{P} \begin{bmatrix} I_{m_1+p} \\ -F_s \end{bmatrix} \begin{bmatrix} \mathcal{A}_L & \widehat{\mathcal{B}} \end{bmatrix},$$

$$\wp = \begin{bmatrix} \mathcal{A}_L & \widehat{\mathcal{B}} \end{bmatrix}^T \begin{bmatrix} I_{m_1+p} \\ -F_s^{\mathcal{P}} \end{bmatrix}^T \mathcal{P} \begin{bmatrix} I_{m_1+p} \\ -F_s^{\mathcal{P}} \end{bmatrix} \begin{bmatrix} \mathcal{A}_L & \widehat{\mathcal{B}} \end{bmatrix},$$

the equality holds if and only if $F_s = F_s^{\mathcal{P}} = -\mathcal{P}_{22}^{-1}\mathcal{P}_{12}^T$. Based on LST, there exists unique $\mathcal{P}$ such that $\gamma \Gamma_{F_s}{}^T \mathcal{P} \Gamma_{F_s} + \Phi_s = \mathcal{P}$. Therefore, we have $J_s^d = J_\Theta$. □

Similarly, there is a strong duality between Primal sub-problem 2b and Dual sub-problem 3b.

Remark 1: Since Lagrange dual function just give a lower bound of the optimal value for primal problem. There exists a gap on the optimal value between primal problem and dual problem in weak duality. The strong duality ensures the equivalence of the two optimal solutions [31]. Therefore, the global optimal solutions of the input constrained sub-problems are obtained by solving their dual sub-problems.

4.3. Sub-optimality for the original system

The decomposability of performance index is discussed in this subsection. It is shown that the optimization of the original problem can be achieved by optimization problems of two subsystems. Accordingly, Algorithm 1 is proposed to learn the composite optimal tracking policy, which is constructed by the tracking policy of slow subsystem and the controller of fast subsystem.

Theorem 2: If there exists a scale $\widetilde{\varepsilon} > 0$, matrices G_0, G_1 and G_2 such that an state feedback tracking policy $u(k) = -G_0 x_1(k) - G_1 x_2(k) - G_2 r(k) + \mathrm{O}(\varepsilon)$ for all $\varepsilon \in (0, \tilde{\varepsilon}]$, then $J_u = J_{u_s} + J_{u_f} + \mathrm{O}(\varepsilon)$, where $G_0 = F_1 - F_f(I - A_4)^{-1}(A_3 - B_2F_1)$, $G_1 = F_f$, $G_2 = (F_f(I - A_4)^{-1}B_2 + I)F_2$, $F_s = \begin{bmatrix} F_1^T & F_2^T \end{bmatrix}^T$.

Proof: Substituting G_0, G_1 and G_2 into $u(k)$, there exists a $\widetilde{\varepsilon} > 0$ to guarantee (38) holds for all $\varepsilon \in (0, \tilde{\varepsilon}]$ [32]

$$u_c(k) = -F_s x_s(k) - F_f x_f(k) + \mathrm{O}(\varepsilon). \tag{38}$$

Obviously, the state feedback tracking policy $u(k)$ consists of three parts

$$u(k) = u_s(k) + u_f(k) + \mathrm{O}(\varepsilon). \tag{39}$$

Therefore,

$$\begin{aligned} J_u &= \sum_{i=k}^{\infty} \gamma^{i-k} \Big[(y_s(k) - r(k))^T Q (y_s(k) - r(k)) \\ &\quad + u_s(k)^T R u_s(k) \Big] + \sum_{i=k}^{\infty} \gamma^{j-k} \Big[y_f(k)^T Q y_s(k) \\ &\quad + u_f(k)^T R u_f(k) \Big] + \mathrm{O}(\varepsilon) \\ &= J_{u_s} + J_{u_f} + \mathrm{O}(\varepsilon) \\ &= J_{u_c} + \mathrm{O}(\varepsilon). \end{aligned} \tag{40}$$

The designed composite control $u(k) = u_s(k) + u_f(k)$ can be regarded as the global optimal solution at $\varepsilon = 0$. Otherwise, it is a sub-optimal controller. The proof of Theorem 2 is completed. □

The strong duality guarantees that the Lagrange dual sub-problems 2a satisfies the following KKT optimality conditions

$$\gamma \Gamma_{F_s} \Theta {\Gamma_{F_s}}^T + \begin{bmatrix} I_{m_1+p} \\ -F_s \end{bmatrix} \Pi \begin{bmatrix} I_{m_1+p} \\ -F_s \end{bmatrix}^T = \Theta, \tag{41}$$

$$\Theta > 0, \tag{42}$$

$$\gamma \Gamma_{F_s}^T P_s \Gamma_{F_s} + \Phi_s - P_s = 0, \tag{43}$$

$$\gamma \left(P_{s_{12}}^T - P_{s_{22}} F_s \right) \begin{bmatrix} \mathcal{A}_L & \widehat{\mathcal{B}} \end{bmatrix} \Theta \begin{bmatrix} \mathcal{A}_L & \widehat{\mathcal{B}} \end{bmatrix}^T = 0. \tag{44}$$

Equation (41) is initial constraints for Primal sub-problem 2a; (42) as a condition of complementary slackness; (43) and (44) can be used to derive policy evaluation and policy improvement. Model-free RL is adopted to solve the problem because $\mathcal{A}_L$ and $\widehat{\mathcal{B}}$ are unknown. Let $\Theta(F_s^{j_s}) = \Theta_{j_s}$, then pre-multiplying equality (43) by $\Theta_{j_s}^T$ and post-multiplying (43) by Θ_{j_s}, we have

$$\Theta_{j_s}^T P_s^{j_s} \Theta_{j_s} = \Theta_{j_s}^T \Phi_s \Theta_{j_s} + \gamma \phi_{j_s}^T P_s^{j_s} \phi_{j_s}, \tag{45}$$

where $\phi_{j_s} = \sum_{k=1}^{\infty} \gamma^{k-1} \begin{bmatrix} \xi(k+1) \\ -F_s^{j_s} \xi(k+1) \end{bmatrix} \begin{bmatrix} \xi(k) \\ -F_s^{j_s} \xi(k) \end{bmatrix}^T$. We can express the kernel matrix $P_s^{j_s}$ as

$$P_s^{j_s} = \Phi_s + \gamma (\phi_{j_s} \Theta_{j_s}^{-1})^T P_s^{j_s} (\phi_{j_s} \Theta_{j_s}^{-1}). \tag{46}$$

There is no need Hamiltonian corresponding Bellman equation to derive the policy gain improvement, the gain can be obtained using condition (44)

$${F_s}^{j_s+1} = \left(P_{s_{22}}^{j_s} \right)^{-1} \left(P_{s_{12}}^{j_s} \right)^T. \tag{47}$$

Let the augmented system $\varsigma_f(k+1) = \Gamma_{F_f} \varsigma_f(k)$, where $\Gamma_{F_f} = \begin{bmatrix} \mathcal{A}_f & \mathcal{B}_f \\ -F_f \mathcal{A}_f & -F_f \mathcal{B}_f \end{bmatrix}$. Applying Online LS method to calculate matrix $P_f^{j_f+1}$ from Bellman equation

$$\left(p_f^{j_f+1} \right)^T \zeta(k) = \vartheta(k) + \gamma \left(p_f^{j_f+1} \right)^T \zeta(k+1), \tag{48}$$

Algorithm 1: The composite optimal algorithm.

Step 1 (Data collection)**:** Collect trajectory date of augmented system $(x_1(k),\ r(k))$.

Step 2 (Initialization)**:** Select stabilizing control policies $u_s^0(k) = -F_s^{j_s}\, \xi(k) + e_s(k)$, $u_f^0(k) = -F_f^{j_f}\, x_f(k)$, where $e_s(k)$ is the exploration noise. ϖ_s and ϖ_f arbitrary small positive scalar. j_s and j_f stand for iteration step staring form 0, the discount factor $0 < \gamma \leq 1$.

Step 3 (Policy evaluation)**:** Action-value function is calculated by solve for matrices $P_s^{j_s+1}$ and $P_f^{j_f+1}$ from the equation

$$\begin{aligned} P_s^{j_s} &= \gamma (\phi_{j_s} \Theta_{j_s}^{-1})^T P_s^{j_s} (\phi_{j_s} \Theta_{j_s}^{-1}) + \Phi_s, \\ {P_f}^{j_f+1} &= \gamma \Gamma_{F_f}^T {P_f}^{j_f+1} \Gamma_{F_f} + \Phi_f. \end{aligned}$$

Step 4 (Policy improvement)**:** Learned policy

$$\begin{aligned} u_s^{j_s+1}(k) &= -(P_{s_{22}}^{j_s+1})^{-1} \left(P_{s_{12}}^{j_s+1} \right)^T \xi(k), \\ u_f^{j_f+1}(k) &= -\left(P_{f_{22}}^{j_f+1} \right)^{-1} \left(P_{f_{12}}^{j_f+1} \right)^T x_f(k). \end{aligned}$$

Step 5 (Termination)**:** Let $j_s = j_s + 1$, $j_f = j_f + 1$ and go to step 3, until $\| P_s^{j_s} - P_s^{j_s+1} \| \leq \varpi_s$ and $\| P_f^{j_f} - P_f^{j_f+1} \| \leq \varpi_f$.

where $\zeta(k) = \varsigma(k) \otimes \varsigma(k)$, $p_f^{j_f+1} = vec(P_f^{j_f+1})$, $\vartheta(k) = \varsigma(k)^T \Phi_f \varsigma(k)$. The policy iteration method based on Lagrangain dual theory is proposed to find the global optimal controller, which makes full use of the knowledge of the fast subsystem.

The composite optimal Algorithm 1 is given.

5. CONVERGENCE AND STABILITY ANALYSIS

In this section, the convergence of policies learned from Algorithm 1 is analyzed theoretically. Then, we adopt the lifting technique and inductive reasoning to prove that the tracking error system is asymptotically stable under the composite sub-optimal tracking policy.

Assumption 2: $(\mathcal{A}_L, \hat{\mathcal{B}})$ and $(\mathcal{A}_f, \mathcal{B}_f)$ are stabilization. $(\mathcal{A}_L, w_s)$ and $(\mathcal{A}_f, w_f)$ are detectable, where $\Phi_s = w_s^T w_s$, $\Phi_f = w_f^T w_f$.

Theorem 3: 1) The tracking policy learning in Algorithm 1 converge to the global optimal solution. 2) The policy $u_s^{j_s+1}$ at every iteration guarantees the asymptotic stability of the tracking error system.

Proof: 1) The convergence of tracking policy can be formulated as $\lim_{j_s \to \infty} P_s^{j_s} = P^*$, $\lim_{j_f \to \infty} P_f^{j_f} = P^{**}$. Taking $\lim_{j_s \to \infty} P_s^{j_s} = P^*$ as an example, note that for any matrix $P_s^{j_s}$, we have

$$P_s^{j_s} = \Gamma_{F_s^{j_s}}^T P_s^{j_s} \Gamma_{F_s^{j_s}} + \Phi_s. \tag{49}$$

By LST, constraint $F_s^{j_s} \in \mathcal{F}_s$ means that the equality (49) has a unique solution $P_s^{j_s} \geqslant 0$. Putting $\Gamma_{F_s^{j_s}} = \begin{bmatrix} \mathcal{A}_L & \widehat{\mathcal{B}} \\ -F_s^{j_s}\mathcal{A}_L & -F_s^{j_s}\widehat{\mathcal{B}} \end{bmatrix}$ in (49) and using Schur Complement, one has

$$P_s^{j_s} \geq \Gamma_{F_s^{j_s+1}}^T P_s^{j_s} \Gamma_{F_s^{j_s+1}} + \Phi_s. \tag{50}$$

Now, defining a mapping according to the inequality (50)

$$\Omega : P_s^{j_s} \to \Gamma_{F_s^{j_s+1}}^T P_s^{j_s} \Gamma_{F_s^{j_s+1}} + \Phi_s, \tag{51}$$

the mapping is non-negative and monotone decreasing. Since $\Phi_s \geqslant 0$ ensures that $\Omega\left(P_s^{j_s}\right) \geqslant 0$ holds for ($j_s = 0$, $1, \cdots$). Moreover, if $P_s^{j_s} \geqslant \widehat{P}_s^{j_s}$, then $\Omega\left(P_s^{j_s}\right) \geqslant \Omega\left(\widehat{P}_s^{j_s}\right)$.

Based on the monotone bounded convergence theorem, there exists $P_s^{j_s+1}$ such that

$$\lim_{l\to\infty} \Omega^l(P_s^{j_s}) = P_s^{j_s+1}. \tag{52}$$

Thus, we can get the $P_s^{j_s} \geq P_s^{j_s+1}$ from $\Omega\left(P_s^{j_s+1}\right) = P_s^{j_s+1}$. Obviously, $\left\{P_s^{j_s}\right\}_{j_s=0,\,1,\,\cdots}$ is non-negative and monotone decreasing, using the monotone bounded convergence theorem again, there exists a matrix $\hat{P}_s$ such that

$$\lim_{j_s\to\infty} P_s^{j_s} = \hat{P}_s. \tag{53}$$

Next, we just need to prove $\hat{P}_s = P^*$. Substituting (11) in the augmented system (10), one can write closed-loop system as

$$\begin{aligned} \xi(k+1) &= \left(\mathcal{A}_L - \widehat{\mathcal{B}}F_s\right)\xi(k), \\ y_s(k) &= \widehat{\mathcal{C}}\xi(k) - \mathcal{D}F_s\xi(k). \end{aligned} \tag{54}$$

The algebraic Riccati equation (ARE) can be expressed as

$$\Upsilon = \hat{Q}_s + \gamma\mathcal{A}_L{}^T\Upsilon A_L - \iota\left(\hat{R}_s + \gamma\widehat{\mathcal{B}}^T\Upsilon\hat{B}\right)^{-1}\iota^T, \tag{55}$$

where $\iota = \left(\tilde{Q}_s + \gamma\mathcal{A}_L{}^T\Upsilon\hat{\mathcal{B}}\right)$. We note that $\Upsilon = \Upsilon^*$ is the unique solution of (55).

There exists a relationship between P^* and the unique solution Υ^* of ARE for closed-loop system (55)

$$P_s^* = \begin{bmatrix} \hat{Q}_s + \gamma\mathcal{A}_L{}^T\Upsilon^*A_L & \tilde{Q}_s + \gamma\mathcal{A}_L{}^T\Upsilon^*\hat{\mathcal{B}} \\ * & \hat{R}_s + \gamma\widehat{\mathcal{B}}^T\Upsilon^*\hat{B} \end{bmatrix}. \tag{56}$$

Combining (49) and (53) results in

$$\Omega\left(\hat{P}_s\right) = \hat{P}_s = \Gamma_{\hat{F}_s}^T\hat{P}_s\Gamma_{\hat{F}_s} + \Phi_s. \tag{57}$$

To show $\hat{P}_s = P^*$, it remains to prove that $\hat{\Upsilon}$ corresponding $\hat{P}_s$ is a solution of ARE (55). Substituting (44) in (57) and noting

$$\hat{\Upsilon} = \hat{P}_{s_{11}} - \hat{P}_{s_{12}}\hat{P}_{s_{22}}^{-1}\hat{P}_{s_{12}}^T, \tag{58}$$

one has

$$\left[\mathcal{A}_L \;\; \widehat{\mathcal{B}}\right]^T\Upsilon\left[\mathcal{A}_L \;\; \widehat{\mathcal{B}}\right] + \Phi_s = P_s, \tag{59}$$

putting $\hat{P}_s = \begin{bmatrix} \hat{Q}_s + \gamma\mathcal{A}_L{}^T\hat{\Upsilon}A_L & \tilde{Q}_s + \gamma\mathcal{A}_L{}^T\hat{\Upsilon}\hat{\mathcal{B}} \\ * & \hat{R}_s + \gamma\widehat{\mathcal{B}}^T\hat{\Upsilon}\hat{B} \end{bmatrix}$ in (59), we can find the result that $\hat{\Upsilon}$ satisfies ARE (55). It indicates $\hat{P}_s = P^*$.

2) We will show that the control policy at the (j_s+1)th ensures the tracking error system asymptotic stability, if a control policy at the j_sth guaranteed the asymptotic stability of the tracking error system. Note tracking error system as

$$e_s(n) = y_s(n) - r(n) = \begin{bmatrix} \mathcal{C} & -I & \mathcal{D} \end{bmatrix}\varsigma_s(k), \tag{60}$$

where $\begin{bmatrix} \mathcal{C} & -I & \mathcal{D} \end{bmatrix} \neq 0$. The system (60) goes to zero asymptotically can be analyzed through the asymptotic stability of $\varsigma_s(n)$, where $\varsigma_s(n)$ is the augmented system (12) measured in slow time scale n . The relationship between the fast time scale k and the slow time scale n is defined by $k = [1/\varepsilon]n = Nn$. Based on lifting technology, Zero-Order Holder and down sampler are employed to obtain the augmented system $\varsigma_s(n)$.

$$\varsigma_s(n+1) = \Gamma_{F_s}\varsigma_s(n), \tag{61}$$

where $\bar{\Gamma}_{F_s} = \begin{bmatrix} \bar{\mathcal{A}}_L & \bar{\mathcal{B}} \\ -F_s\bar{\mathcal{A}}_L & -F_s\bar{\mathcal{B}} \end{bmatrix}$, $\bar{\mathcal{A}}_L = \begin{bmatrix} \mathcal{A}_s^N & 0 \\ 0 & L^N \end{bmatrix}$, $\bar{\mathcal{B}} = \begin{bmatrix} \sum_{g=0}^{N-1}\mathcal{A}_s\mathcal{B}_s \\ 0 \end{bmatrix}$. We have $F_s^{j_s} \in \mathcal{F}_s$ according to the assumption of asymptotical stability at the j_sth, and further demonstrate on $F_s^{j_s+1} \in \mathcal{F}_s$, where

$$\mathcal{F}_s = \left\{F_s : \rho\left(\bar{\mathcal{A}}_L - \bar{\mathcal{B}}F_s\right) < 1\right\}. \tag{62}$$

By lemma LST, there exists a gain matrix $W = \begin{bmatrix} \lambda_1^{\frac{1}{2}}Q_1 & \lambda_2^{\frac{1}{2}}Q_2 \\ \lambda_1^{\frac{1}{2}}Q_2^T & \lambda_2^{\frac{1}{2}}Q_3 \end{bmatrix}$ such that $\rho\left(\mathcal{A}_L + KW\right) < 1$. Let $Z = \begin{bmatrix} \lambda_1^{\frac{1}{2}} & \left(KQ_2 + \lambda_2^{-\frac{1}{2}}\hat{\mathcal{B}}\right)Q_3^{-1} \\ \frac{1}{2}F_s^{j_s+1}\hat{\mathcal{B}} & F_s^{j_s+1}\hat{\mathcal{B}}V^{-1} \end{bmatrix}\begin{bmatrix} \lambda_1^{\frac{1}{2}}Q_1 & \lambda_2^{\frac{1}{2}}Q_2 \\ \lambda_1^{\frac{1}{2}}Q_2^T & \lambda_2^{\frac{1}{2}}Q_3 \end{bmatrix}$, inequation $\rho(\Gamma_{F_s^{j_s+1}} + Zw) < 1$ holds, that is, $\left(\Gamma_{F_s^{j_s+1}},\, W\right)$ is detectable. Using LST again, $F_s^{j_s+1} \in \mathcal{F}_s$ is derived, control gain at (j_s+1)th guarantees the asymptotic stability of the tracking error system. □

Remark 2: The fast subsystem is asymptotically stable under the global optimal policy. In addition, the control policies u_s and u_f are independent, the composite sub-optimal tracking policy makes the original system asymptotically stable [31].

6. SIMULATION EXAMPLE

In this section, a F-8 aircraft with two time-scales is taken as an experimental example to proved that the Q-learning framework approach performs satisfactorily [33]. In performing Algorithm 1, the initial stabilizing policies are selected according to data of the original system and the reference system. In Subsection 6.1, the formulations of F-8 aircraft and tracking goal are given. The simulation results of the proposed method are presented in Subsection 6.2. In Subsection 6.3, we verify that Algorithm 1 is effective and then prove the superiority in the tracking performance and practicability by comparing with existing methods [6,11,14].

6.1. The dynamic formulation of F-8 aircraft and tracking goal

The longitudinal dynamics of F-8 aircraft is given by (63).

$$\begin{bmatrix}\dot{\upsilon}(t)\\ \dot{\kappa}(t)\\ \dot{\alpha}(t)\\ \dot{q}(t)\end{bmatrix} = G\begin{bmatrix}\upsilon(t)\\ \kappa(t)\\ \alpha(t)\\ q(t)\end{bmatrix} + B\delta_e(t),$$
$$\begin{bmatrix}n_z(t)\\ q(t)\end{bmatrix} = C\begin{bmatrix}\upsilon(t)\\ \kappa(t)\\ \alpha(t)\\ q(t)\end{bmatrix}, \tag{63}$$

where $G=\begin{bmatrix} X_\upsilon & -g/V_0 & X_\alpha/V_0 & 0\\ 0 & 0 & 0 & 1\\ Z_\upsilon V_0 & 0 & Z_\alpha & 1\\ M_\upsilon V_0 & 0 & M_\alpha & M_q\end{bmatrix}$, $B=\begin{bmatrix}X_{\delta_e}/V_0\\ 0\\ Z_{\delta_e}\\ M_{\delta_e}\end{bmatrix}$, $C=\begin{bmatrix}0 & dM_\kappa - Z_\kappa \acute{V} & dM_\alpha - Z_\alpha \acute{V} & dM_q\\ 0 & 0 & 0 & 1\end{bmatrix}$. Parameters in Table 1 are partially known,where Z_{δ_e}, M_{δ_e}, X_υ are unavailable.

Parameters in F-8 aircraft and its physical meaning.

Table 1. Parameters in F-8 aircraft and its physical meaning.

Parameter	Physical meaning
V_0	Total equilibrium velocity
$M_{()}$	Stable axes stability derivatives
$X_{()}, Z_{()}$	Wind axes stability derivatives
q	Pitch rate
κ	Incremental pitch attitude
α	Incremental angle of attack
δ_e	Incremental elevator position
V	Equilibrium velocity
g	Acceleration due to gravity
n_z	Normal acceleration
d	Accelerometer displacement
$\acute{V}$	Airstream velocity

According to [34], the two-time-scale property are demonstrated by a proper scaling. Rewriting the matrices of the state dynamics (63) in the form $G=\begin{bmatrix}\varepsilon G_{11} & G_{12}\\ \varepsilon G_{21} & G_{22}\end{bmatrix} = \begin{bmatrix}0 & G_{12}\\ 0 & G_{22}\end{bmatrix} + \varepsilon\begin{bmatrix}0 & G_{11}\\ 0 & G_{21}\end{bmatrix} = G_0 + \varepsilon G_1$, and a standard singularly perturbed system is thus obtained

$$\begin{bmatrix}\dot{x}_1\\ \varepsilon\dot{x}_2\end{bmatrix} = \begin{bmatrix}A_1 & A_2\\ A_3 & A_4\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} + \begin{bmatrix}B_1\\ B_2\end{bmatrix}u, \tag{64}$$

where x_1, x_2 and u are mentioned in Subsection 2.2, $A_1=\tilde{\mathrm{H}}G_1\tilde{\mathrm{Z}}$, $A_2=\tilde{\mathrm{H}}G_1\tilde{\mathrm{Z}}$, $A_3=\tilde{\mathrm{H}}G_1\tilde{\mathrm{Z}}$, $A_4=\tilde{\mathrm{H}}G_0\tilde{\mathrm{Z}}+\varepsilon\tilde{\mathrm{H}}G_1\tilde{\mathrm{Z}}$, $B_2=\begin{bmatrix}Z_{\delta_e}\\ M_{\delta_e}\end{bmatrix}$, $B_1=\begin{bmatrix}\frac{X_{\delta_e}}{\varepsilon V_0}\\ 0\end{bmatrix} - G_{12}G_{22}^{-1}B_2$, $\tilde{\mathrm{H}}=\begin{bmatrix}I & -G_{12}G_{22}^{-1}\\ 0 & I\end{bmatrix}$, $\tilde{H}=\begin{bmatrix}0\\ I\end{bmatrix}^T$, $\begin{bmatrix}\bar{\mathrm{H}}\\ \tilde{H}\end{bmatrix}^{-1} = \begin{bmatrix}\bar{\mathrm{Z}}^T\\ \tilde{\mathrm{Z}}^T\end{bmatrix}^T$. After discretization, the F-8 aircraft model is presented as the following discrete-time singularly perturbed dynamics (1). Obviously, the slow state vector and the fast state vector have different response speeds in singularly perturbed system.

The command generator dynamics is described by

$$r(k+1) = \begin{bmatrix}-0.740 & -0.200\\ -0.200 & -0.900\end{bmatrix} r(k). \tag{65}$$

We collect a set of initial values $r(0)=\begin{bmatrix}-1 & 0.3\end{bmatrix}$, $X(0)=[x_1(0)\ \ x_2(0)]=\begin{bmatrix}0.1 & 0.5 & -1 & -0.5\end{bmatrix}^T$. The discount factor and weighting matrices are selected as $\gamma=0.9$, $Q=10$, $R_s=R_f=0.1$.

6.2. Simulation results of the proposed method

In this subsection, the simulation results of the F-8 aircraft are presented by using Algorithm 1. The selection of the stabilizing policy depends on $r(0)$ and $X(0)$.

Trajectories of the slow subsystem, the desired goal and the tracking error system are displayed in Figs. 3 and 4. Fig. 5 shows that iteration steps of the global optimal tracking gain F_s^* and the global optimal control gain F_f^*, respectively. The optimal feedback gains are reached respectively at 10 and 43 iterations, and are denoted as $F_s^*=\begin{bmatrix}-0.0285 & -0.6551 & 0.4135 & 0.1822\end{bmatrix}$, $F_f^*=\begin{bmatrix}-3.7834 & -4.8132\end{bmatrix}$. In addition, the controller u_f and the tracker u_s account for the composite tracker u_c in Fig. 5. Response of the fast subsystem is presented in Fig. 6. Exploration noise in Fig. 6 is used to inspire potential information of system in the model-free Q-learning framework method. The amplitude of the reverse M-sequence is small due to value of the optimal tracking policy, and the maximum is selected no more than fifteen percent.

6.3. Comparison with existing methods

In this subsection, ARE method, linear MPC method and model-free RL method are adopted to compare with the proposed method.

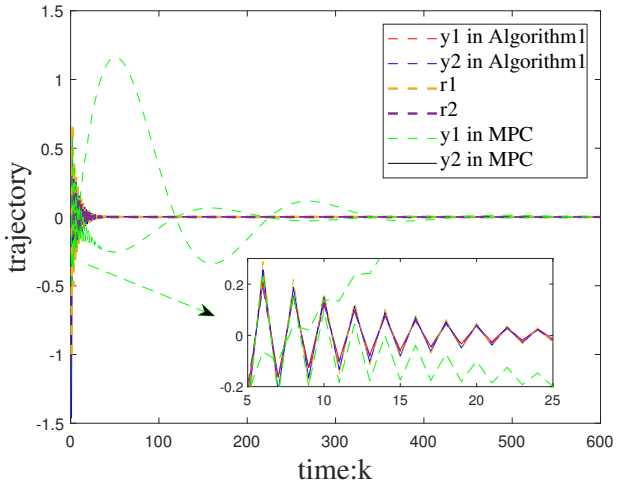

Fig. 3. Trajectories of the system and desired goal.

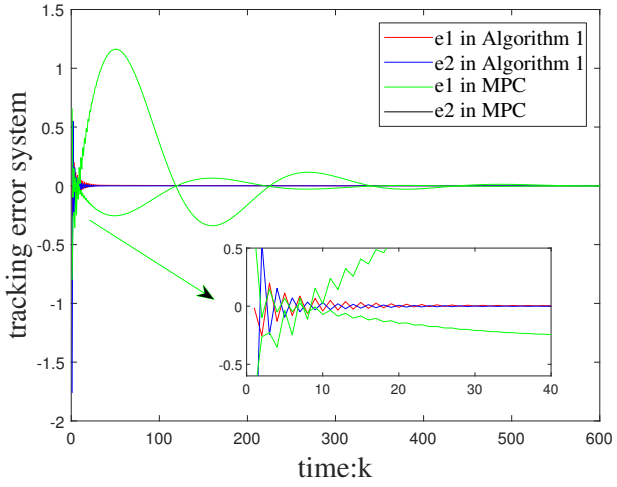

Fig. 4. Fast subsystem and excitation signal.

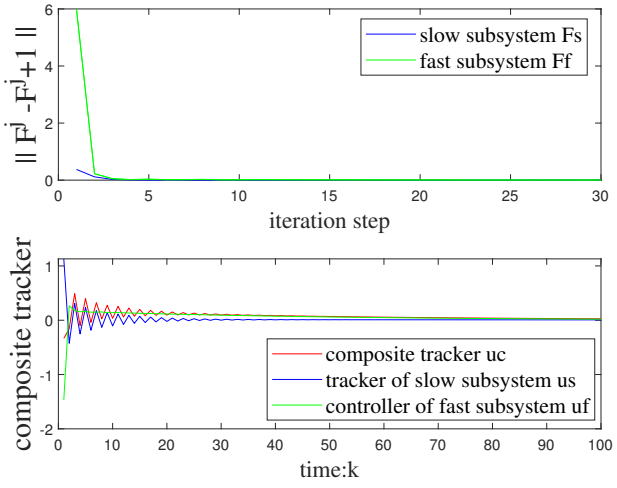

Fig. 5. Iteration step and composite tracker.

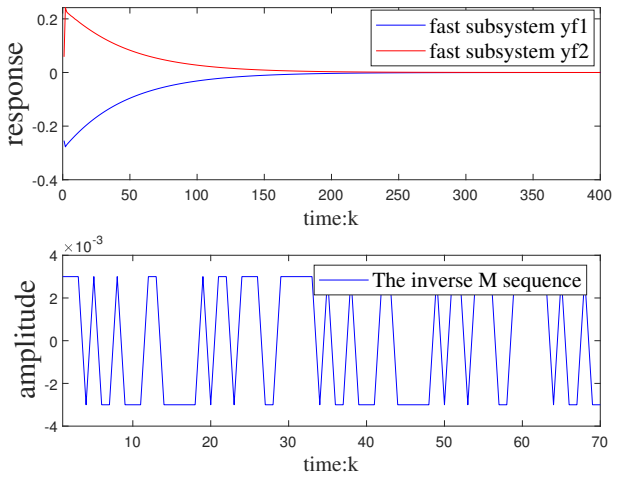

Fig. 6. Iteration step and composite tracker.

Table 2. Comparison of the optimal performance index.

Method	Performance index
The proposed method	$J_c^* = 40.6976$
ARE method in [9]	$J_{opt} = 40.6970$

Assuming that the model information is complete, then the comparison of the optimal performance index is displayed in Table 2. The learning cost loss with model information partially unknown is approximately 0.05%, which proved that the proposed method is effective. It is point that the proposed method reduces the dependence on model.

Compared with the linear MPC method [6], the tracking performance of the proposed method is better. The parameters of model are obtained by parameter identification, where $I+\varepsilon A_1 = \begin{bmatrix} 0.9930 & -0.0227 \\ 0.0495 & 0.9995 \end{bmatrix}$, $\varepsilon A_2 = \begin{bmatrix} -0.0307 & 0.0034 \\ -0.0007 & 0.0001 \end{bmatrix}$, $\varepsilon B_1 = \begin{bmatrix} 0.0048 \\ -0.5693 \end{bmatrix}$. The parameters $B_2 = \begin{bmatrix} -0.0022 \\ -0.1139 \end{bmatrix}$, $A_3 = \begin{bmatrix} -0.0015 & 0 \\ 0.0005 & 0 \end{bmatrix}$, $A_4 = \begin{bmatrix} 0.9886 & 0.0130 \\ -0.0625 & 0.9932 \end{bmatrix}$ are known. Discount factor $\gamma = 0.9$. Set prediction horizon and control horizon as 8.

Trajectories of original system and tracking error system are presented in Figs. 3 and 4. As can be seen from Figs. 3 and 4, amplitude and convergence rate of tracking error system and trajectories of system obtained by the linear MPC method are at a distinct disadvantage, which may be caused by ill-condition and variable coupling. As shown in Table 3, integral absolute error (IAE) and mean square error (MSE) of the proposed method are smaller. Therefore, the proposed method is superior to the linear MPC method in tracking performance and convergence speed for the TTS systems.

The practicability of the proposed Algorithm 1 is im-

Table 3. Comparison of error.

$k^* = 500, n = 200$	IAE	MSE
Proposed method	0.0201	2.0212e-4
MPC method in [6]	0.6593	0.0047

proved in contrast to Algorithm 2 [12]. In Algorithm 2, the initial stabilizing policy in RL method is described as an admissible policy. However, the literature provided no concrete standard to select the admissible policy when the model information is unknown. Based on convex optimization, we present the standard by considering the admissible policy as an input constraint. The selections of the admissible policy only need to satisfy (21) and (27).

7. CONCLUSION

A sub-optimal tracking control of discrete-time TTS system with partially unknown dynamics is discussed in this paper. The initial stabilizing policy is considered as an input constraint and solved by using LST and LMI. Two reduced-order input constrained optimization problems corresponding to the sub-optimal tracking control problem are proposed to improve tracking performance. Then, based on convex optimization and RL technique, the Q-learning framework method and policy iteration method are employed to solve the reduced-order problems. A global sub-optimal composite policy is obtained. We also analyze the asymptotic stability of tracking error system.

CONFLICTS OF INTERESTS

The authors declare that there is no competing financial interest or personal relationship that could have appeared to influence the work reported in this paper.

REFERENCES

[1] A. Raza, F. M. Malik, N. Mazhar, and R. Khan, "Two-time-scale robust output feedback control for aircraft longitudinal dynamics via sliding mode control and high-gain observer," *Alexandria Engineering Journal*, vol. 61, no. 6, pp. 4573-4583, October 2022.

[2] N. Daroogheh, N. Meskin, and K. Khorasani, "Ensemble kalman filters for state estimation and prediction of two-time scale nonlinear systems with application to gas turbine engines," *IEEE Transactions on Control Systems Technology*, vol. 27, no. 6, pp. 2565-2573, September 2019.

[3] J. Yang, P. Si, Z. Wang, X. Jiang, and L. Hanzo, "Dynamic resource allocation and layer selection for scalable video streaming in femtocell networks: A twin-time-scale approach," *IEEE Transactions on Communications*, vol. 66, no. 8, pp. 3455-3470, August 2018.

[4] J. Kim, U. Jon, and H. Lee. "State-constrained sub-optimal tracking controller for continuous-time linear time-invariant (CT-LTI) systems and its application for DC motor servo systems," *Applied Sciences*, vol. 10, no. 16, pp. 5724-5741, August 2020.

[5] G. B. Avanzini, A. Zanhettin, and P. Rocco, "Constrained model predictive control for mobile robotic manipulators," *Robotica*, vol. 36, no. 1, pp. 19-38, April 2018.

[6] V. R. Saksena, J. Oreilly, and P. V. Kokotovic, "Singular perturbations and time-scale methods in control theory: Survey 1976-1983," *Automatica*, vol. 20, no. 3, pp. 273-293, May 1984.

[7] V. Dragan. "On the linear quadratic optimal control for systems described by singularly perturbed it differential equations with two fast time scales," *Axioms*, vol. 8, no. 1, pp. 1-30, March 2019.

[8] W. Chen, Y. Liu, and W. X. Zheng, "Synchronization analysis of two-time-scale nonlinear complex networks with time-scale-dependent coupling," *IEEE Transactions on Cybernetics*, vol. 49, no. 9, pp. 3255-3267, September 2019.

[9] W. Xue, J. Fan, V. G. Lopez, J. Li, Y. Jiang, T. Chai, and F. L. Lewis, "New Methods for Optimal Operational Control of Industrial Processes Using Reinforcement Learning on Two Time Scales," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 5, pp. 3085-3099, May 2020.

[10] W. Xue, J. Fan, V. G. Lopez, Y. Jiang, T. Chai, and F. L. Lewis, "Off-Policy Reinforcement Learning for Tracking in Continuous-Time Systems on Two Time Scales," *IEEE Transactions on Neural Networks and Learning System*, vol. 32, no. 10, pp. 4334-4346, October 2021.

[11] R. Sutton, A. Barto, *Reinforcement Learning - An Introduction*, MIT Press, Cambridge, 1998.

[12] X. Wu and C. Wang, "Model-free optimal tracking control for an aircraft skin inspection robot with constrained-input and input time-delay via integral reinforcement learning," *International Journal of Control, Automation, and Systems*, vol. 18, pp. 245-257, January 2020.

[13] Y. Yang, K. G. Vamvoudakis, H. Modares, Y. Yin, and D. C. Wunsch, "Hamiltonian-driven hybrid adaptive dynamic programming," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 10, pp. 6423-6434, October 2021.

[14] T. Lindner, A. Milecki, and D. Wyrwa, "Positioning of the robotic arm using different reinforcement learning algorithms," *International Journal of Control, Automation, and Systems*, vol. 19, pp. 1661-1676, April 2021.

[15] V. Vu, Q. Tran, T. Pham, and P. N. Dao, "Online actor-critic reinforcement learning control for uncertain surface vessel systems with external disturbances," *International Journal of Control, Automation, and Systems*, vol. 20, pp. 1029-1040, March 2022.

[16] Y. Peng, Q. Chen, and W. Sun, "Reinforcement Q-learning algorithm for H infinite tracking control of unknown discrete-time linear systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 11, pp. 4109-4122, November 2020.

[17] L. Zhou, J. Zhao, L. Ma, and C. Yang, "Decentralized composite suboptimal control for a class of two-time-scale interconnected networks with unknown slow dynamics," *Neurocomputing*, vol. 383, no. 21, pp. 71-79, March 2020.

[18] M. Sayak, B. He, and C. Aranya, "Reduced-dimensional reinforcement learning control using singular perturbation approximations," *Automatica*, vol. 126, no. 21, pp. 1-11, April 2021.

[19] K. Bahare, F. L. Lewis, M. Hamidreza, A. Karimpour, and M.-B. Naghibi-Sistani, "Reinforcement Q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics," *Automatica*, vol. 50, no. 4, pp. 1167-1175, April 2014.

[20] Y. Jiang, J. Fan, T. Chai, F. L. Lewis, and J. Li, "Tracking control for linear discrete-time networked control systems with unknown dynamics and dropout," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 10, pp. 4607-4620, October. 2018.

[21] S. A. A. Rizvi, A. J. Pertzborn, and Z. Lin, "Reinforcement learning based optimal tracking control under unmeasurable disturbances with application to HVAC systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 15, no. 4, pp. 1-11, June 2021.

[22] X. F. Li, L. Xue, and C. Y. Sun, "Linear quadratic tracking control of unknown discrete-time systems using value iteration algorithm," *Neurocomputing*, vol. 314, no. 7, pp. 86-93, November 2018.

[23] Y. Jiang, J. Fan, T. Chai, and F. L. Lewis, "Dual-rate operational optimal control for flotation industrial process with unknown operational model," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 6, pp. 4587-4599, June 2019.

[24] J. Li, B. Kiumarsi, T. Chai, F. L. Lewis, and J. Fan, "Off-policy reinforcement learning: optimal operational control for two-time-scale industrial processes," *IEEE Transactions on Cybernetics*, vol. 47, no. 12, pp. 4547-4558, December 2017.

[25] G. Gu, *Discrete-time Linear Systems: Theory and Design with Applications*, Springer, New York, NY, USA, 2012.

[26] P. Kokotovic, H. K. Khalil, and J. Oreilly, *Singular Perturbation Methods in Control: Analysis and Design, Society for Industrial and Mathematics*, Philadelphia, PA, 1999.

[27] V. Mayuresh, "Robust constrained model predictive control using linear matrix inequalities," *Automatica*, vol. 32, no. 10, pp. 1361-1379, February 1996.

[28] K. R. Muske, "Model predictive contro with linear models," *AIChE Journal*, vol. 49, no. 9, pp. 3255-3267, September 1993.

[29] D. Lee and J. Hu, "Primal-dual Q-learning framework for LQR design," *IEEE Transactions on Automatic Control*, vol. 64, no. 9, pp. 3756-3763, September 2019.

[30] F. Zhang, *The Schur Complement and Its Applications*, vol. 4, Springer, New York, NY, USA, 2006.

[31] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[32] B. Litkouhi and H. Khalil, "Multirate and composite control of two-time-scale discrete-time systems," *IEEE Transactions on Automatic Control*, vol. 30, no. 7, pp. 645-651, July 1985.

[33] J. Elliott, "NASA's advanced control law program for the F-8 digital fly-by-wire aircraft," *IEEE Transactions on Automatic Control*, vol. 22, no. 5, pp. 753-757, October 1977.

[34] P. V. Kokotovi, *Singular Perturbation Methods in Control: Analysis and Design*, London, 1986.

Xuejie Que received her M.S. degree in applied mathematics from the Zhengzhou University, Zhengzhou, China, in 2019. She is currently pursuing a Ph.D. degree in Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai, China. Her research interests include multi-time-scale systems, optimal control, and reinforcement learning.

Zhenlei Wang is currently a Professor with the School of Information Science and Engineering, East China University of Science and Technology and also with the Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education. His research interests include intelligent control, modeling and analysis the characteristic of complex systems, intelligent optimization algorithms, and fault diagnosis.

Xin Wang is currently an Associate Professor in Shanghai Jiao Tong University, China. His research interests include multi-variable intelligent decoupling control, control and optimization of complex industrial processes, and multiple model adaptive control.