Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

Reinforcement Learning-Assisted Composite Adaptive Control for Time-Varying Parameters

Seong-hun Kim * **Hanna Lee** ** **Youdan Kim** ***

* *Department of Aerospace Engineering, Seoul National University, Seoul, Korea, (e-mail: bgbgof@snu.ac.kr).*
** *Department of Aerospace Engineering, Seoul National University, Seoul, Korea, (e-mail: hn.lee@snu.ac.kr).*
*** *Department of Aerospace Engineering, Seoul National University, Seoul, Korea, (e-mail: ydkim@snu.ac.kr)*

Abstract: Adaptive control methods have received a lot of interest to control uncertain systems with parametric uncertainties. In particular, composite adaptation law that incorporates a memory storing the past trajectory data is promising, because it has an exponential convergent rate for both the tracking error and the parameter estimation under a mild condition of excitation. In this study, this research direction is extended to cope with uncertain parameters that change over time, which is difficult to solve with traditional memory-based methods. The problem is formulated into a Markov decision process, and a reinforcement learning algorithm is adopted to solve the optimal decision making problem. The proposed formulation preserves the stability of the original composite adaptive system, and the reinforcement learning agent can learn the optimal composite strategy.

Keywords: Model reference adaptive control, Intelligent control, Learning control, Uncertain dynamic systems, Markov decision processes

1. INTRODUCTION

Adaptation of a controller to regulate the systems under uncertainties in the mathematical models has been the main principle of direct adaptive control schemes. Model reference adaptive control (MRAC), one of the most popular methods, can ensure the Lyapunov stability for a certain class of systems which track a pre-designed reference model. This is why MRAC schemes have been widely used in the safety-critical systems such as robotics and aerospace applications.

However, there is one disadvantage in implementing the MRAC schemes for physical systems. Even if the Lyapunov stability analysis guarantees the convergence of the target states, the responses in transient stages may be either too rapid for the physical controller to generate such control inputs, or undesirably very slow. This trade-off is determined by the adaptation gain in update law of the control parameters, which corresponds to the learning rate in the context of online learning. Several studies have been conducted to alleviate this problem by combining uncertainty estimators. Slotine and Li (1991) suggested a composite MRAC which adds a parameter estimation error to the update equation, in a way that it does not compromise stability. They expected that more accurate estimates of the uncertain parameters would enhance the transient responses, which has been supported by many case studies, such as Lavretsky (2009), Nakanishi et al. (2005), and Patre et al. (2010).

Given a dataset, the accuracy of estimation depends on the quality of the dataset, which can be assessed by its spectral characteristics. In the context of the composite MRAC, however, data is revealed at each time following a certain trajectory in a high-dimensional space. Thus, it is inevitable to introduce a new tool to measure the quality of the incoming data, which is called the *persistent excitation* condition (Anderson (1977), and Boyd and Sastry (1986)). If this condition is met, the estimation of uncertain parameters becomes more accurate, which implies that the composite MRAC provides better performance. However, the persistent excitation sometimes requires a perturbation in the input command to the system, which is undesirable for some real applications such as a level-flight. Chowdhary et al. (2014) resolved this issue by considering that the data observed in the past also contains useful information if the uncertain parameters remain constant. The concurrent-learning adaptive control they introduced reuses a set of data stored in a memory, and updates this memory concurrently with observations in real-time. Cho et al. (2018), Pan and Yu (2016), and Pan and Yu (2018) suggested additional continuous-time systems that can replace the discrete-time memory to analyze the behavior more apparently.

⋆ This work was supported by the advanced Research Center Program(NRF-2019R1A2C2083946) through the National Research Foundation of Korea(NRF) grant funded by the Korean Government(MSIP) contracted through the Institute of Advanced Aerospace Technology at Seoul National University.

The common concept of these approaches is to find the best combination of the data previously observed to maximize a given performance index. This concept is very sound to improve data efficiency in dynamic systems operating online, but there exist two limitations. First, the optimization relies only on the observed data. Indeed, the optimization results affect subsequent observations, because they are included in the parameter update equation. Second, the optimization is sensitive to time-varying parameters. It is hard to capture the true parameters through this kind of data processing, when the parameters are varying over time.

To solve the above-mentioned problems, in this study, a novel memory-based composite MRAC method is proposed, which is assisted by reinforcement learning (RL) techniques. Because the memory consists of the past data experienced by the system, it is essentially a non-Markovian process if the RL agent chooses action based on it. Thus, state is redefined to formulate the whole system into the Markov decision process (MDP). The RL agent determines the importance of each past data stored in the memory, and this decision is reflected in the adaptation law. The importance of each data is assessed as expected total reward, which includes tracking errors and control efforts as well as the traditional memory quality assessment such as the minimum eigenvalue. Thanks to the original structure of the composite MRAC, the uncertain probabilistic decisions of the RL agent can be readily applied to the adaptation law without compromising stability. The online implementation of reinforcement learning also ensures that the direction of parameter update is optimal in the long term sense.

This paper is organized as follows. Section 2 provides a brief introduction to a system class considered in this study, and a basic structure of composite adaptive controller along with the stability analysis. Section 3 demonstrates a memory-driven composite MRAC scheme which preserves the stability. Section 4 introduces an MDP formulation of the memory-driven model, and a RL algorithm to solve the problem, appropriately. The result of numerical simulation is presented in Section 5. Concluding remark is given in Section 6.

2. PRELIMINARIES

2.1 Notation

Let $\mathbb{R}$ and $\mathbb{N}$ denote the set of real numbers and natural numbers, respectively. Also, let $\mathbb{R}^d$ denote the set of all real vectors in d dimensions. Given a real vector v, let $\|v\| = \sqrt{v^T v}$ denote the Euclidean norm, while given a real matrix Q, the Frobenius norm $\|Q\| = \sqrt{\mathrm{tr}(Q^T Q)}$ is used instead, where $\mathrm{tr}(\cdot)$ denotes the trace operator. For a given square matrix R, let $\lambda_{\min}(R)$ and $\lambda_{\max}(R)$ denote the minimum and maximum eigenvalue, respectively.

2.2 Problem Formulation

The MRAC designs require a main model and a bounded-input-bounded-output (BIBO) stable reference model. Commands are inserted to the reference model, and appropriate feedback controllers are designed so that the main model follows the reference model. Let us define an error as the difference between their states, and regard the error dynamics as our main system to be analyzed. Even if the error dynamics depends on the states of the main and reference models, it can be reflected in time-dependent functions of a non-autonomous system as follows,

$$\dot{x} = f(t,x) + B(u + \delta(t,x)), \tag{1}$$

where $x \in \mathcal{X}$ is the state, $u \in \mathcal{U}$ is the control input, the functions f and δ are piecewise continuous in t and locally Lipschitz in x on $[0,\infty) \times \mathcal{X}$, and the constant matrix B has full column rank. The function f and the matrix B are assumed to be known, while the function δ represents all uncertainties in the model. The uncertainty δ considered in this paper is represented as follows,

$$\delta(t,x) = W^\circ(t)^T \phi(x), \tag{2}$$

where $\phi : \mathcal{X} \to \mathbb{R}^d$ is the known basis function, and the function W° denotes the unknown time-varying parameter. The function ϕ is locally Lipschitz in x, and the norms $\|W^\circ\|$ and $\|\dot{W}^\circ(t)\|$ are bounded for all $t \geq 0$.

Let us further assume that $x = 0$ is an exponentially stable equilibrium point of the following system.

$$\dot{x} = f(t,x). \tag{3}$$

This is a typical assumption for the error dynamics in the literature of MRAC methods. The Lyapunov converse theorem (Khalil (2002)) implies that there exists a function V° such that

$$V^\circ(t,x) \geq c_1 \|x\|^2, \tag{4}$$

$$\frac{\partial V^\circ}{\partial t} + \frac{\partial V^\circ}{\partial x} f(t,x) \leq -c_2 \|x\|^2, \tag{5}$$

for all $[0,\infty) \times \mathcal{X}$, where c_1 and c_2 are positive scalars.

2.3 Composite MRAC

Let the control input be

$$u = -W^T \phi, \tag{6}$$

where W is the control parameter updated by the following adaptation law:

$$\dot{W} = \Gamma \phi(x) \frac{\partial V^\circ}{\partial x} B - \Gamma(F(t)W - G(t)), \tag{7}$$

where the constant matrix Γ is positive definite, and the matrix-valued functions F and G are piecewise continuous in $t \geq 0$.

Theorem 1. Suppose that the matrix $F(t)$ is positive definite, and the norm $\|F(t)W^\circ(t) - G(t)\|$ is bounded for all $t \geq 0$. Then, with the control input (6), the solution (x, W) to the systems (1) and (7) is uniformly ultimately bounded.

Proof. Suppose that there exist positive scalars b_1 and b_2 such that

$$\|F(t)W^\circ(t) - G(t)\| \leq b_1,$$
$$\|\dot{W}^\circ(t)\| \leq b_2,$$

for all $t \geq 0$. Let the Lyapunov function candidate be

$$V = V^\circ + \frac{1}{2} \mathrm{tr}\left(\tilde{W}^T \Gamma^{-1} \tilde{W}\right), \tag{8}$$

where $\tilde{W} \coloneqq W - W^\circ$. The time derivative of V is given by

$$\begin{aligned}\dot{V} &\leq -c_2\|x\|^2 - \operatorname{tr}\left(\tilde{W}^T(FW-G)\right) \\ &\quad - \operatorname{tr}\left(\tilde{W}^T\Gamma^{-1}\dot{W}^\circ\right) \\ &\leq -c_2\|x\|^2 - \lambda_{\min}(F)\|\tilde{W}\|^2 \\ &\quad - \operatorname{tr}\left(\tilde{W}^T(FW^\circ-G)\right) - \operatorname{tr}\left(\tilde{W}^T\Gamma^{-1}\dot{W}^\circ\right) \\ &\leq -c_2\|x\|^2 - \lambda_{\min}(F)\|\tilde{W}\|^2 \\ &\quad + \left(b_1 + \lambda_{\min}(\Gamma)^{-1}b_2\right)\|\tilde{W}\|.\end{aligned}$$

The rest of the proof is similar to that in Chowdhary et al. (2014).

3. MEMORY-DRIVEN COMPOSITE MRAC

The overall stability of the composite MRAC system can directly be guaranteed by Theorem 1, where the conditions for F and G can be satisfied by various methods (Chowdhary et al. (2014) and Cho et al. (2018)). This section provides a simple approach to configure the functions F and G based on a memory that stores the observed data of filtered signals. An update law for the memory is then proposed. Along with the overall MRAC system, the update law can be expressed as an MDP, as discussed in the next section, which implies that any RL algorithms enable optimal construction of the memory.

3.1 Stability Analysis

To satisfy the condition of Theorem 1 uniformly in time, consider the following filtered system (Cho et al. (2018)).

$$\dot{\varphi} = -\frac{1}{\tau}(\varphi - \phi), \tag{9}$$

$$\dot{z} = \frac{1}{\tau}(B^\dagger x - z) + B^\dagger f(t,x) + u, \tag{10}$$

$$y = \frac{1}{\tau}(B^\dagger x - z), \tag{11}$$

where $B^\dagger \coloneqq (B^TB)^{-1}B^T$, and the positive scalar τ is a time constant. Let $\epsilon = y - W^{\circ T}\varphi$, then,

$$\dot{\epsilon} = -\frac{1}{\tau}\epsilon - \dot{W}^{\circ T}\varphi. \tag{12}$$

Since x and $\|\dot{W}^\circ\|$ are bounded, ϕ, φ and ϵ are also bounded, which implies the following linearly parameterized, time-varying relation.

$$y(t) = W^{\circ T}(t)\varphi(t) + \epsilon(t). \tag{13}$$

Proposition 2. Let

$$F(t) = \varepsilon(t)I_d + \sum_i a_i\varphi(t_i)\varphi(t_i)^T, \tag{14}$$

$$G(t) = \sum_i a_i\varphi(t_i)y(t_i)^T, \tag{15}$$

for a finite sequence of time $\{t_i\}$, and of positive bounded real numbers $\{a_i\}$, where $I_d \in \mathbb{R}^{d\times d}$ is an identity matrix. The function $\varepsilon : [0,\infty) \to [0,\varepsilon_0]$ denotes an auxiliary function to ensure $\lambda_{\min}(F(t)) > 0$ for all $t \geq 0$. Then, the conditions of Theorem 1 are satisfied with (14) and (15).

Proof. The matrix $F(t)$ is obviously positive definite. Now, suppose that there exist positive scalars b_3 and b_4 such that

$$\|W^\circ(t)\| \leq b_3, \quad \|\epsilon(t)\| \leq b_4, \tag{16}$$

for all $t \geq 0$. Then, from (13), (14), and (15), we have

$$\begin{aligned}&\|F(t)W^\circ(t) - G(t)\| \\ &\quad \leq \varepsilon_0 b_3 + \left\|\sum_i a_i\varphi(t_i)\varphi(t_i)^T(W^\circ(t) - W^\circ(t_i))\right\| \\ &\qquad + \left\|\sum_i a_i\varphi(t_i)\epsilon(t_i)^T\right\| \\ &\quad \leq \varepsilon_0 b_3 + \sum_i a_i\|\varphi(t_i)\|\left(2b_3\|\varphi(t_i)\| + b_4\right).\end{aligned}$$

Since φ is bounded, and the sequences $\{t_i\}$ and $\{a_i\}$ are finite, the norm $\|F(t)W^\circ(t) - G(t)\|$ is also bounded for all $t \geq 0$.

Note that Proposition 2 implies the system (1) is uniformly ultimately bounded with F and G being any finite sums of past data with finite weights in the form of (14) and (15). A straight forward choice of the auxiliary function is

$$\varepsilon(t) = \varepsilon_0\mathbb{1}_{\{\lambda_{\min}(F^\circ(t))=0\}}, \tag{17}$$

for some $\epsilon_0 > 0$, where $F^\circ(t) \coloneqq \sum_i a_i\varphi(t_i)\varphi(t_i)^T$, and $\mathbb{1}_{\{\cdot\}}$ denotes an indicator function.

Now, what remains is to select appropriate sequence of weights, a_i, based on the sequential observations. The next section will show that the problem can be formulated as an MDP with a proper selection of an extended state space.

3.2 Memory Update Law

To implement RL algorithms, discrete-time update laws are selected for F and G. The RL agent observes state information at each time $t \in \mathcal{T} \coloneqq \{t_i\}$, where $i \in \mathbb{N}$, and $t_1 = 0$. Hence, it is required that F and G defined in (14) and (15) are constructed by the observed data.

For $n \in \mathbb{N}$, let us define an index set as

$$\mathcal{I}_n \coloneqq \left\{1 \leq i_1^{(n)} < \ldots < i_{\nu(\mathcal{I}_n)}^{(n)} = n\right\}, \tag{18}$$

where $\nu(\mathcal{I}_n) \leq N \in \mathbb{N}$ denotes the size of the index set, and a corresponding sequence of non-negative scalars as

$$a^{(n)} \coloneqq (a_1^{(n)}, \ldots, a_{\nu(\mathcal{I}_n)}^{(n)}). \tag{19}$$

Given $\mathcal{T}$, let

$$F(t) = F_n \coloneqq \varepsilon(t_n)I_d + \sum_{i\in\mathcal{I}_n} a_i^{(n)}\hat{\varphi}(t_i)\hat{\varphi}(t_i)^T, \tag{20}$$

$$G(t) = G_n \coloneqq \sum_{i\in\mathcal{I}_n} a_i^{(n)}\hat{\varphi}(t_i)\hat{y}(t_i)^T, \tag{21}$$

for all $t \in [t_n, t_{n+1})$, where $t_n, t_{n+1} \in \mathcal{T}$, the function $\hat{\varphi}(\cdot) \coloneqq \varphi(\cdot)/\|\varphi(\cdot)\|$, the function $\hat{y}(\cdot) \coloneqq y(\cdot)/\|\varphi(\cdot)\|$, and the function ε is the auxiliary function. The functions F and G defined above satisfy the conditions of Proposition 2. Note that

$$\varepsilon(t_n) \leq \lambda_{\min}(F_n) \leq \lambda_{\max}(F_n) \leq \varepsilon(t_n) + \sum_{i\in\mathcal{I}_n} a_i^{(n)}, \tag{22}$$

by Weyl's inequality.

Unlike the previous methods in Pan and Yu (2018) and Cho et al. (2018), the algorithm proposed in this study actively chooses each $a_i^{(n)}$ at each stage n based on a set of data $\mathcal{D}_n$, called a memory, denoted by

$$\mathcal{D}_n \coloneqq \{(\hat{\varphi}(t_i), \hat{y}(t_i))\}_{i\in\mathcal{I}_n}. \tag{23}$$

Equations (20) and (21) can be rewritten with $\mathcal{D}_n$ as

$$F_n = F_n(a^{(n)}, \mathcal{D}_n), \quad G_n = G_n(a^{(n)}, \mathcal{D}_n), \tag{24}$$

with slight abuses of notation.

To keep the size of $\mathcal{D}_n$ finite, a purging algorithm is developed based on $a^{(n)}$. Let

$$i_{\min}^{(n)} := \arg\min_{i \in \mathcal{I}_n} a_i^{(n)}. \tag{25}$$

The index set $\mathcal{I}_n$ is updated as follows,

$$\mathcal{I}_{n+1} = \begin{cases} \mathcal{I}_n \cup \{n+1\}, & \text{if } \nu(\mathcal{I}_n) < N, \\ \mathcal{I}_n \setminus \{i_{\min}^{(n)}\} \cup \{n+1\}, & \text{else.} \end{cases} \tag{26}$$

The role of the index $i_{\min}^{(n)}$ is to pull out the most irrelevant data in the memory $\mathcal{D}_n$, which is also updated according to (26). Note that the memory always contains the most recent observation.

4. RL-ASSISTED COMPOSITE MRAC

This section provides an MDP formulation for the suggested data-driven MRAC. When the problem is properly formulated into an MDP, it is straightforward to implement any RL algorithms.

4.1 MDP Formulation

Let us rewrite the system dynamics into a typical but abstract MRAC form to see that the overall systems can be represented as an MDP. The main model, reference model, and command model are given by

$$x(t_{n+1}) \sim p_x(X \mid x(t_n), x_r(t_n), W(t_n)), \tag{27}$$
$$x_r(t_{n+1}) = f_r(x_r(t_n), c(t_n)), \tag{28}$$
$$c(t_{n+1}) \sim p_c(c \mid c(t_n)), \tag{29}$$

where x and x_r are the state vectors of the main and reference models, respectively, and c is the command which is assumed to have an unknown dynamics. The stochastic property, represented by the probabilities p_x and p_c, may come from unobservable states such as W°, and from the internal dynamics of c. The update laws for W and $\mathcal{D}_n$ can be represented as

$$W(t_{n+1}) = f_W(x(t_n), x_r(t_n), W(t_n), \mathcal{D}_n, a^{(n)}), \tag{30}$$
$$\mathcal{D}_{n+1} \sim p_{\mathcal{D}}(\mathcal{D} \mid x(t_n), x_r(t_n), \mathcal{D}_n, a^{(n)}), \tag{31}$$

where the probability $p_{\mathcal{D}}$ may originate in the differences between the dynamics of x and x_r, and of the filtered values φ and y.

From the above observations, if an extended state is defined by

$$s^{(n)} := (x(t_n), x_r(t_n), c(t_n), W(t_n), \mathcal{D}_n), \tag{32}$$

it is obvious that the probability of the next state $s^{(n+1)} \in \mathcal{S}$ is determined only by the current state $s^{(n)} \in \mathcal{S}$ and the current action $a^{(n)} \in \mathcal{A}$, which is defined in (19), i.e., $s^{(n+1)} \sim p(s \mid s^{(n)}, a^{(n)})$. The overall structure of the RL-assisted composite MRAC scheme is depicted in Fig. 1.

To formulate an MDP for RL, it is necessary to define a reward, which has a finite value at each stage. Consider the following reward.

$$r(s, a) := \beta_1(\lambda_{\min}(F)) - \beta_2(\|x\|) - \beta_3(\|u\|), \tag{33}$$

for $(s, a) \in \mathcal{S} \times \mathcal{A}$, where u is the control input defined in (6), $\beta_1, \beta_2, \beta_3 : [0, \infty) \to [0, \infty)$ are strictly increasing functions, and $\|\cdot\|$ denotes any norm. Since F, x, and u are bounded, the reward at each stage is also finite. MDP is now well defined by the tuple $(\mathcal{S}, \mathcal{A}, p, r)$.

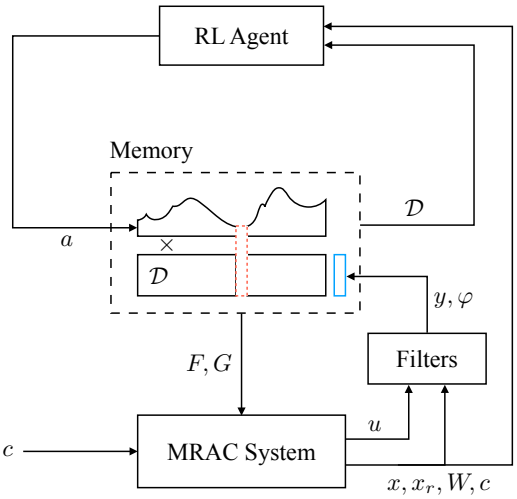

Fig. 1. The block diagram of the RL-assisted composite MRAC. The blue rectangle denotes most recently observed data appended to the memory, where the red-dotted rectangle denotes pulled out data.

4.2 RL Algorithm

The MDP formulation proposed in Section 4.1 requires a model-free RL algorithm that runs in real-time. Running in real-time necessitates off-policy algorithms to reuse past experience. Deep deterministic policy gradient (DDPG) is the commonly used algorithm for such settings with continuous state and action spaces (Lillicrap et al. (2016)). The algorithm provides for sample efficiency, but is easily unstable and highly sensitive to hyperparameters. Haarnoja et al. (2018) proposed an efficient but stable model-free, off-policy RL algorithm, called soft-actor-critic (SAC). In this framework, the actor tries to maximize an expected reward while maximizing entropy. As SAC has shown promising results in complex domains, such as robotics, the algorithm is implemented for the proposed RL-assisted composite MRAC framework.

5. NUMERICAL EXAMPLE

Numerical simulation is conducted to demonstrate the performance of the proposed method compared to the existing CMRAC methods. The model used in the simulation is a virtual-control-augmented model of wing rock for slender delta wing. The uncertain parameters are 1,000-times larger than the original parameters from Elzebda et al. (1989), and bounded continuous functions of time are added to change the parameters over time. The model is augmented to perform a tracking control for a given reference model, and a corresponding feedback controller is implemented according to the design guide in Cho et al. (2018). For the convenience, only the tracking error model and the reference model are shown here as

$$\dot{e} = \begin{bmatrix} 0 & 1 & 0 \\ -15.8 & -5.6 & -17.3 \\ 1 & 0 & 0 \end{bmatrix} e + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \left(u + W^{\circ}(t)^T \phi(t,e)\right), \tag{34}$$

$$\dot{x}_r = \begin{bmatrix} 0 & 1 & 0 \\ -15.8 & -5.6 & -17.3 \\ 1 & 0 & 0 \end{bmatrix} x_r - \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} c(t), \tag{35}$$

where the function c is piecewise continuous on $[0, \infty)$. The time-varying parameter is defined by

$$W^{\circ}(t) = \begin{bmatrix} -18.59521 \\ 15.162375 \\ -62.45153 \\ 9.54708 \\ 21.45291 \end{bmatrix} + \begin{bmatrix} 30 \\ 30 \\ 30 \\ 30 \\ 30 \end{bmatrix} \tanh\left(\frac{t}{60}\right), \tag{36}$$

and the basis function is given as

$$\phi(t,e) = \left[x_1, x_2, |x_1|x_2, |x_2|x_2, x_1^3\right]^T, \tag{37}$$

where $x := e + x_r(t)$. The function c denotes a command signal for x_1 to be tracked. The initial states are all set to be zero except for $x(0) = [0.3, 0, 0]^T$. The time constant for the filtered systems (9), (10), and (11) is chosen to be 0.001. The memory size is chosen as $N = 100$, and the observation occurs in every 0.01 second.

The hyperparameters for SAC is given in Table 1. The notations are the same as Haarnoja et al. (2018). The value network, the soft-Q network, and the policy network have three linear hidden layers with ReLU activation functions. The reward function is chosen as (33), where $\beta_1 = 10^5, \beta_2 = 100$, and $\beta_3 = 0.1$.

The results of the proposed method, referred to as RL-CMRAC, are compared to the standard MRAC, and the composite MRAC proposed in Cho et al. (2018), referred to as FE-CMRAC. The full state and control histories of each method are shown in Fig. 2. The reference model tracking error seems to be small enough in the sense of average, while there exist high-frequency responses for the standard MRAC and FE-CMRAC. Particularly noteworthy, the high-frequency oscillation gradually disappears in the standard MRAC, while it is growing in the FE-CMRAC. This phenomena is mainly due to the time-varying feature of the uncertain parameters. The FE-CMRAC still seems to determine that the past data are more important, where the parameters are quite different from the current.

Figure 3 supports this interpretation. The memory of the FE-CMRAC is no longer updated after about 34 seconds. In contrast, the RL-CMRAC actively updates its memory. In spite of the fluctuation, the minimum eigenvalue of F tends to increase gradually, and after 90 seconds the RL-CMRAC outperforms the FE-CMRAC.

Finally, the parameter estimation errors are depicted in Fig. 4, with the real parameters. It can be seen that the estimation error goes to zero for the RL-CMRAC, while the FE-CMRAC does not.

Table 1. Hyperparamters for SAC.

Hyperparameter	Value	Hyperparameter	Value
γ	0.99	$\lambda_V, \lambda_Q, \lambda_\pi$	10^{-4}
τ	0.01	Batch size	128
Buffer size	10^4	Hidden layer size	32

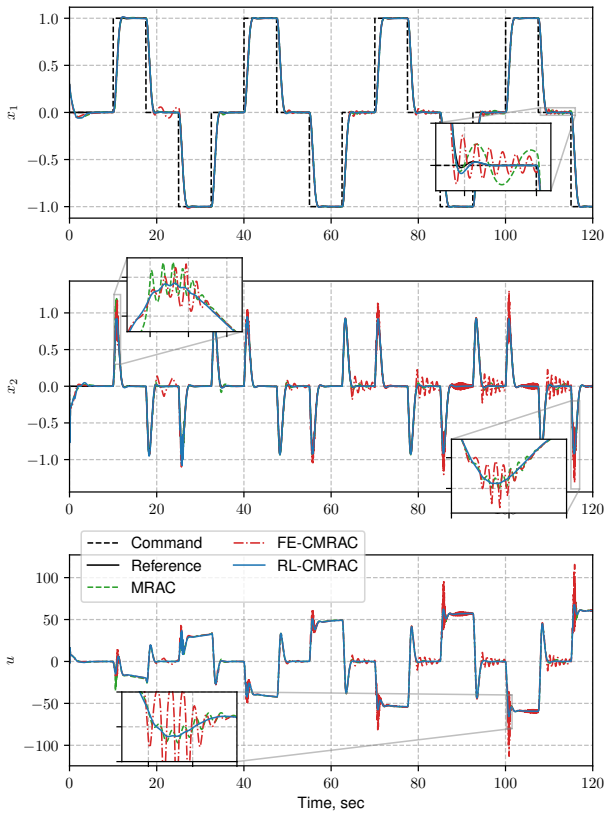

Fig. 2. State and control input history.

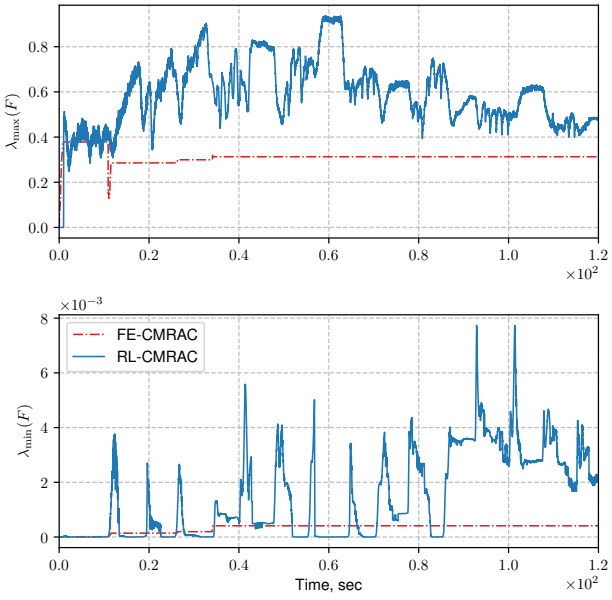

Fig. 3. Maximum (top) and minimum (bottom) eigenvalue history.

6. CONCLUSION

A new RL-assisted composite MRAC scheme was proposed to overcome the essential limitations of the existing memory-based methods. For this purpose, an MDP using a novel state representation was formulated considering the memory update structure. The probabilistic RL-based

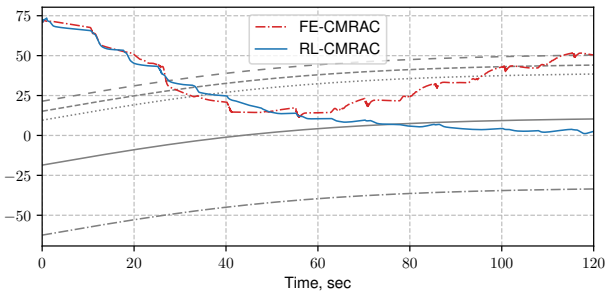

Fig. 4. Parameter estimation error ($\|\tilde{W}\|$) history. Gray lines are the real parameters.

controller was carefully selected and implemented, which preserves the original stability of the composite MRAC methods. Numerical simulation result confirmed the convergence of both tracking error and parameter estimation error even if the parameter changes over time.

REFERENCES

Anderson, B. (1977). Exponential stability of linear equations arising in adaptive identification. *IEEE Transactions on Automatic Control*, 22(1), 83–88.

Boyd, S. and Sastry, S. (1986). Necessary and sufficient conditions for parameter convergence in adaptive control. *Automatica*, 22(6), 629–639.

Cho, N., Shin, H., Kim, Y., and Tsourdos, A. (2018). Composite model reference adaptive control with parameter convergence under finite excitation. *IEEE Transactions on Automatic Control*, 63(3), 811–818.

Chowdhary, G., Mühlegg, M., and Johnson, E. (2014). Exponential parameter and tracking error convergence guarantees for adaptive controllers without persistency of excitation. *International Journal of Control*, 87(8), 1583–1603.

Elzebda, J., Nayfeh, A., and Mook, D. (1989). Development of an analytical model of wing rock for slender delta wings. *Journal of Aircraft*, 26(8), 737–743.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*. Stockholm, Sweden.

Khalil, H. (2002). *Nonlinear Systems*. Prentice Hall, Upper Saddle River, NJ.

Lavretsky, E. (2009). Combined/composite model reference adaptive control. *IEEE Transactions on Automatic Control*, 54(11), 2692–2697.

Lillicrap, T., Hunt, J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2016). Continuous control with deep reinforcement learning. In *4th International Conference on Learning Representations (ICLR)*. San Juan, Puerto Rico.

Nakanishi, J., Farrell, J., and Schaal, S. (2005). Composite adaptive control with locally weighted statistical learning. *Neural Networks*, 18(1), 71–90.

Pan, Y. and Yu, H. (2016). Composite learning from adaptive dynamic surface control. *IEEE Transactions on Automatic Control*, 61(9), 2603–2609.

Pan, Y. and Yu, H. (2018). Composite learning robot control with guaranteed parameter convergence. *Automatica*, 89, 398–406.

Patre, P., MacKunis, W., Johnson, M., and Dixon, W. (2010). Composite adaptive control for Euler–Lagrange systems with additive disturbances. *Automatica*, 46(1), 140–147.

Slotine, J. and Li, W. (1991). *Applied Nonlinear Control*. Prentice Hall, Englewood Cliffs, NJ.