Learning Tracking Control for Cyber-Physical Systems

Wu, Chengwei; Pan, Wei; Sun, Guanghui; Liu, Jianxing; Wu, Ligang

DOI
10.1109/JIOT.2021.3056633

Publication date
2021

Document Version
Accepted author manuscript

Published in
IEEE Internet of Things Journal

Learning Tracking Control for Cyber-Physical Systems

Chengwei Wu, Wei Pan, Guanghui Sun, Jianxing Liu, and Ligang Wu, *Fellow, IEEE*

***Abstract*—This paper investigates the problem of optimal tracking control for cyber-physical systems (CPS) when the cyber realm is attacked by denial-of-service (DoS) attacks which can prevent the control signal transmitting to the actuator. Attention is focused on how to design the optimal tracking control scheme without using the system dynamics and analyze the impact of DoS attacks on tracking performance. First, a Riccati equation for the augmented system including the system model and the reference model is derived under the framework of dynamic programming. The existence and uniqueness of its solution are proved. Second, the impact of the successful DoS attack probability on tracking performance is analyzed. A critical value of the probability is given, beyond which the solution to the Riccati equation cannot converge. The tracking controller cannot be designed. Third, reinforcement learning is introduced to design the optimal tracking control schemes, in which the system dynamics are not necessary to be known. Finally, both a dc motor and an F16 aircraft are used to evaluate the proposed control schemes in this paper.**

***Index Terms*—Cyber-physical systems, Reinforcement learning, Optimal tracking control, DoS attacks.**

1.. INTRODUCTION

The increasing development of computer and communication devices promotes the emergence and application of cyber-physical systems (CPS). The cyber realm ubiquitously embeds such devices to process, exchange and gather the information and then directly interacts with the physical components. As a promising engineered system, it can be applied to a variety of fields varying from the national defense to smart housing. Especially, the thriving of 5G, which provides a more reliable and low-delay communication network, will make CPS more and more widely applied in the future. Antsaklis in [1] has scrutinized the relevant definition, applications and challenges of CPS and pointed out that CPS would transform the way that human interacts with the physical environment. It is worth noting that CPS bring great advantages while challenges cannot be neglected due to the vulnerability of the cyber layer. Most researchers motivate their research by discussing some cyber attacks. Examples of such attacks include both Stuxnet and attacking Maroochy Shire Council's sewage control systems [2]. It has been an urgent task to design schemes to secure CPS against attacks.

To address the security problem, many researchers have dedicated to such a field [3]–[8]. For the secure state estimation problem, the critical condition to securely reconstruct the state under sparse sensor attacks has been proposed in [9]. Combining the sliding mode observer, a secure algorithm has been proposed to reconstruct states under both sparse attacks and external disturbances in [2]. To make the schemes in [2], [9] more robust to attacks, a secure estimation reconstruction algorithm, which allows the attacks to change over time has been provided in [10]. Besides the secure state reconstruction, the secure control under attacks also attracted considerable attention. In [11], an adaptive control framework has been proposed to mitigate the sensor and actuator false data injection attacks. To reduce the assumptions imposed on the denial-of-service (DoS) attack model, attack frequency and attack duration approaches have been proposed in [12]. For the secure consensus control for multi-agent systems under attacks, it can refer to [13], [14] and the references therein. Different from the above results, the game-theoretical approach, which can address an attacker and a defender in a unified framework has been applied to design secure defense control schemes [15]–[17]. Nevertheless, the secure tracking control problem has not been fully investigated except [18]. Besides, the exact system dynamics are the necessary knowledge in the aforementioned results.

Reinforcement learning can find optimal decisions without using exact system dynamics. Such a technique refers to two different forms. One is that the underlying environment is described by a Markov decision process. In this scenario, the reinforcement learning is often used in games and robotic control [19]. But the stability of the learned policies is not guaranteed. The other is that the environment is described by differential equations without knowing exact system dynamics. Using the model structure, the second reinforcement learning technique can not only design optimal control policies, but also guarantee the stability, which has been widely applied in the control field [20]–[22]. Combining the Q-learning approach [23], [24], the reinforcement learning approach, also known as the adaptive dynamic programming in the control community has been widely applied to find solutions for different control problems, for example, zero-sum game based optimal control [25], [26], optimal control for linear periodic systems [27], control for networked systems [28], and tracking controller design [29], [30]. Although elegant control schemes have been proposed, there exist two problems in the Q-learning approach [21]. One is that if the external disturbance is considered, it

This work was supported in part by the National Key R&D Program of China (No. 2019YFB1312001), National Natural Science Foundation of China (62033005, 62022030, 62003114), and the State Grid Heilongjiang Electric Power Company Limited funded project (No. 522417190057). *Corresponding author: Ligang Wu*

C. Wu, G. Sun, J. Liu and L. Wu are all with the Department of Control Science and Engineering, Harbin Institute of Technology, Harbin 150001, P.R. China. E-mail: ligangwu@hit.edu.cn

W. Pan is with the Department of Cognitive Robotics, Delft University of Technology, Netherlands.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3056633, IEEE Internet of Things Journal

needs to evolve in a specific manner. The other lies in that the probing noise adding to the control input can result in the bias of the solution. In [21], an off-line optimal controller has been designed, and the mentioned two problems have been solved. It is noted that the secure tracking control problem for CPS using reinforcement learning is not fully studied. Compared with existing results, for example, [21], [30], [31], there exist some challenges in designing the secure tracking controller under attacks. The challenges include analyzing the existence of the solution to the derived Riccati equation, and revealing the relation between the attack probability and the system performance, which motivates this paper.

This paper investigates the secure tracking control problem for CPS under malicious actuator DoS attacks, which are modeled based on the signal-to-interference-plus-noise (SINR) ratio based communication model. The reference model which can be unstable is given to generate the tracked signal. Reinforcement learning is introduced to design the secure tracking controller. The main contributions of this paper can be summarized as follows:

1) This paper shows that the value function for the augmented system (i.e., augmenting the reference model and the physical plant) can be rewritten in the quadratic form, with which the Bellman equation is used to derive the Riccati equation.
2) Different from existing results, for example, [21], [30], the successful attack probability exists in the derived Riccati equation. The probability affects the existence and uniqueness of the solution to the Riccati equation. This paper proves that the existence and uniqueness of the solution to the derived Riccati equation can be guaranteed under certain conditions.
3) A critical condition for the successful attack probability is derived, beyond which the solution to the Riccati equation cannot converge. Using the learning scheme and the matrix decomposition technique, the critical value is obtained without using the exact system dynamics.

Finally, both a dc motor and an F16 aircraft system are utilized to evaluate the effectiveness of the proposed control scheme.

The rest of this paper is organized as follows. Section 2. describes the system formulation and the problem setup. Section 3. introduces how to prove the existence and optimality of the desired tracking controller. A Q-learning based control scheme and its convergence are provided in Section 4.. An off-policy learning control scheme is proposed in Section 5., and then we conclude this paper in Section 6..

Notations. The notations used throughout the paper are defined as follows. $\mathcal{A}^T$ means the transpose of the matrix $\mathcal{A}$. $\mathcal{M}^{-1}$ is the inverse of the matrix $\mathcal{M}$. $\mathbb{R}^n$ denotes the n-dimensional Euclidean space. A positive definite (positive semidefinite) matrix $\mathcal{P}$ is defined as $\mathcal{P} > 0 (\mathcal{P} \geq 0)$. I and 0 represent the identity matrix and a zero matrix with compatible dimensions, respectively. $\text{diag}(\cdot)$ denotes the matrix with diagonal structure. $\otimes$ denotes the Kronecker products and $\text{vec}(\mathcal{Q})$ is a column vector consisting of the transpose of each row in $\mathcal{Q}$. Without explicitly stated, the dimensions of matrices are compatible with algebraic operation.

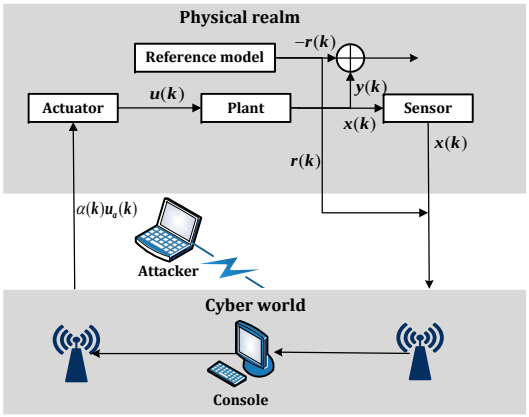

Fig. 1. The system blueprint. The definitions of the symbols $x(k)$, $y(k)$, $u(k)$ etc are defined in the paper.

2.. System Formulation and Preliminaries

The diagram of the system frame is described in Fig. 1, which shows that the controller and the actuator interact with each other using the cyber layer. The adversary can implement DoS attacks to prevent the cyber realm from transmitting the control signal to the actuator. In this section, we give a model to describe the physical dynamics. A command model is provided to generate the reference signal. An SINR-based communication model is introduced to describe the interactions between the system designer and the adversary. Following the above setup, the control objective of this paper is set. Next, we give the details.

A. Physical process and reference model descriptions

In this paper, we assume that the underlying physical plant in Fig. 1 is governed by the following model [32]–[34]

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k), \\ y(k) &= Cx(k), \end{aligned} \tag{1}$$

where $x(k) \in \mathbb{R}^{n_x}$ is the state vector, $u(k) \in \mathbb{R}^{n_u}$ represents the control input signal and $y(k) \in \mathbb{R}^{n_y}$ denotes the measurement output. A, B and C are matrices with appropriate dimensions, which are unknown.

For the reference trajectory, we give the following command generator

$$r(k+1) = Fr(k), \tag{2}$$

where $r(k)$ is the reference trajectory and F means a given gain. Here, F is not necessary to be Hurwitz.

Remark 1: As discussed in [29], such a model can generate a variety of trajectories, for example, the step signal, the ramp and the sinusoidal waveform. As to difficulties resulting from using non-Hurwitz F, it can be overcome by introducing a discount factor in the designed performance index. The details will be discussed later.

Considering system (1) and the reference trajectory (2), the augmented system is written as

$$\bar{x}(k+1) = \bar{A}\bar{x}(k) + \bar{B}u(k), \tag{3}$$

where

$$\bar{x}(k) = \begin{bmatrix} x(k) \\ r(k) \end{bmatrix}, \ \bar{A} = \begin{bmatrix} A & 0 \\ 0 & F \end{bmatrix}, \ \bar{B} = \begin{bmatrix} B \\ 0 \end{bmatrix}.$$

For the tracking controller, it is designed as follows

$$u_a(k) = K_1 x(k) + K_2 r(k), \tag{4}$$

where K_1 and K_2 are tracking controller gains, which are designed in the following contents.

B. SINR-based communication model

For DoS attacks, the attacker can implement them with many alternative techniques. For instance, it can send superfluous requests to flood the targeted communication network to attempt to prevent all legitimate requests from being fulfilled. Also, the system operator has the power to send requests to the communication. Accordingly, to address the interaction between the system operator and the adversary, a signal-to-interference-plus-noise ratio based communication model is introduced to describe the successful attack probability in this section. The relation between the symbol error rate R_{SER} and signal to noise ratio R_{SNR} is described as [35]

$$\begin{aligned} R_{SER} &= 2\mathcal{S}(\sqrt{\varphi R_{SNR}}), \\ \mathcal{S}(\nu) &= \frac{1}{\sqrt{2\pi}} \int_{\nu}^{\infty} e^{\frac{-\xi^2}{2}} d\xi, \end{aligned} \tag{5}$$

where φ is a positive scalar.

Based on the digital communication theory [36], R_{SNR} under DoS attacks can be rewritten as

$$R_{SINR} = \frac{\varrho_s(k)}{\varrho_a(k) + \varsigma^2},$$

where R_{SINR} means the signal-to-interference-plus-noise ratio, $\varrho_s(k)$ and $\varrho_a(k)$ respectively mean the power, which the system operator and the attacker utilize to send their requests at time k. ς^2 is the additive white Gaussian noise power.

If DoS attacks occur, they can result in packet dropouts [12], [34], [37]. To describe the effect of DoS attacks on the actuator, we define $\alpha(k)$ as an indicator function. When $\alpha(k) = 0$, it indicates that the attack is successfully implemented, and the packet is lost. Otherwise, $\alpha(k) = 1$. Then, we can obtain

$$u(k) = \alpha(k) u_a(k), \tag{6}$$

According to the above discussion, the following equation can be obtained [35]

$$\bar{\alpha} = \mathbb{P}(\alpha(k) = 1) = \left(1 - 2\mathcal{S}\left(\sqrt{\frac{\varphi \varrho_s(k)}{\varrho_a(k) + \varsigma^2}}\right)\right)^L, \tag{7}$$

where L means the length of the transmitted data.

C. Control objective

In this paper, not only the tracking performance but also the optimality should be guaranteed. Thus, define the following value function to quantify the control cost

$$\begin{aligned} V(\bar{x}(k)) &= \mathbb{E}\left\{\sum_{i=k}^{\infty} \beta^{i-k} \Big[(y(k) - r(k))^T Q (y(k) - r(k)) \right. \\ &\quad \left. + u^T(k) R u(k) \Big] \right\} \\ &= \mathbb{E}\left\{\sum_{i=k}^{\infty} \beta^{i-k} \left(\bar{x}^T(k) \bar{Q} \bar{x}(k) \right.\right. \\ &\quad \left.\left. + \alpha(k) u_a^T(k) R u_a(k) \right) \right\}, \end{aligned} \tag{8}$$

where $0 < \beta \leq 1$ is defined as a discounted factor, the known weighting matrices $Q \geq 0$, $R > 0$, and $\bar{Q} = \bar{C}^T Q \bar{C}$, $\bar{C} = \begin{bmatrix} C & -I \end{bmatrix}$.

Remark 2: The value function similar to (8) can be found in [29], [30], whose results cannot be extended to our paper directly due to the existence of the stochastic indication function $\alpha(k)$. The expectation of $\alpha(k)$ affects the existence of the tracking controller to be designed, which will be discussed later. Besides, the discussion of the discounted factor β can refer to [29].

In this paper, the control objective is to design a learning based optimal tracking scheme such that the performance index (8) can be minimized (i.e., minimizing the control cost) while the output signal can be driven to track the desired trajectory. Next, we will show how to derive the desired optimal tracking controller.

3.. Optimal tracking controller design and stability analysis

This section mainly presents how to design the optimal tracking controller. First, we prove that the value function is still in the quadratic form even DoS attacks occur. Second, the Riccati equation for the augmented system (3) is derived, by solving which the optimal tracking controller is designed. Last, the existence and uniqueness of the solution to the Riccati equation are proved. The relation between the critical value of the attack probability and the existence and uniqueness of the solution is revealed.

A. Analysis of the value function

First, a proposition is proposed to show that the value function defined in (8) can be written in a quadratic form.

Proposition 1: If the tracking control scheme is designed as $u(k) = \alpha(k) u_a(k)$, the performance index (8) can be written as $V(\bar{x}(k)) = \mathbb{E}\left\{\bar{x}^T(k) P \bar{x}(k)\right\}$ with

$$P = \begin{bmatrix} P_1 & P_2 \\ P_2^T & P_3 \end{bmatrix}.$$

Proof: Submitting the controller $u(k)$ into the performance index (8) yields

$$\begin{aligned} V(\bar{x}(k)) &= \mathbb{E}\left\{\sum_{i=k}^{\infty} \beta^{i-k} \left(\bar{x}^T(k) \bar{C}^T \bar{Q} \bar{C} \bar{x}(k) \right.\right. \\ &\quad \left.\left. + \alpha(k) u_a^T(k) R u_a(k) \right) \right\} \\ &= \mathbb{E}\left\{\sum_{i=0}^{\infty} \beta^i \left[\bar{x}^T(i+k) \left(\bar{Q} + \alpha(k) K_1^T R K_1 \right) \right.\right. \\ &\quad \times \bar{x}(i+k) + \bar{x}^T(i+k) \left(-C^T Q + \alpha(k) \right. \\ &\quad \left.\left. \times K_1^T R K_2 \right) r(i+k) + r^T(i+k) \left(-QC \right. \right. \\ &\quad \left.\left. + \alpha(k) K_2^T R K_1 \right) \bar{x}(i+k) + r^T(i+k) \right. \end{aligned}$$

$$\times \left(Q+\alpha(k)K_2^TRK_2\right)r(i+k)\big]\big\}. \quad (9)$$

Using the dynamics in the augmented system (3) and the command generator (2), $\bar{x}(i+k)$ and $r(i+k)$ can be computed as

$$\begin{aligned}\bar{x}(i+k) &= G_i\bar{x}(k)+H_ir(k),\\ r(i+k) &= F^ir(k),\end{aligned} \quad (10)$$

where

$$h_j = A+\alpha(k+j)BK_1,\ G_i=\prod_{j=0}^{i}h(j),$$

$$H_i = \sum_{n=0}^{i-1}\left(\prod_{j=n}^{i-n-1}h(j+1)\right)\alpha(k+n)BK_2F^n,$$

for the operator '$\prod$', if the upper bound is less than or equal to the lower bound, $h(\bullet)=1$.

Combining (9) and (10) yields

$$V(\bar{x}(k)) = \mathbb{E}\left\{\bar{x}^T(k)P\bar{x}(k)\right\},$$

where

$$\begin{aligned}P_1 &= \sum_{i=0}^{\infty}\beta^i\left[G_i^T\left(\bar{Q}+\alpha(k)K_1^TRK_1\right)G_i\right],\\ P_2 &= \sum_{i=0}^{\infty}\beta^i\big[G_i^T\left(-C^TQ+\alpha(k)K_1^TRK_2\right)F^i\\ &\quad+G_i^T\left(\bar{Q}+\alpha(k)K_1^TRK_1\right)H_i\big],\\ P_3 &= \sum_{i=0}^{\infty}\beta^i\big[(F^i)^T\left(Q+\alpha(k)K_2^TRK_2\right)F^i\\ &\quad+H_i^T\left(\bar{Q}+\alpha(k)K_1^TRK_1\right)H_i\\ &\quad+H_i^T\left(-C^TQ+\alpha(k)K_1^TRK_2\right)F^i\\ &\quad+(F^i)^T\left(-QC+\alpha(k)K_2^TRK_1\right)H_i\big].\end{aligned}$$

The proof is completed. ■

B. Optimal tracking controller design

Next, a theorem is proposed to determine the gains K_1 and K_2 in the tracking controller.

Theorem 1: For the system (1), the optimal tracking controller is designed as

$$u_a(k) = K_1x(k)+K_2r(k) = -\bar{K}\bar{x}(k), \quad (11)$$

where $\bar{K} = \left(R+\beta\bar{B}^TP\bar{B}\right)^{-1}\bar{B}^TP\bar{A}$ and P is a unique solution to the following Riccati equation

$$\begin{aligned}P &= \bar{Q}+\beta\bar{A}^TP\bar{A}\\ &\quad-\bar{\alpha}\beta^2\bar{A}^TP\bar{B}\left(R+\beta\bar{B}^TP\bar{B}\right)^{-1}\bar{B}^TP\bar{A}.\end{aligned} \quad (12)$$

Proof: According to the performance index (8) and $V(\bar{x}(k))=\mathbb{E}\left\{\bar{x}^T(k)P\bar{x}(k)\right\}$, we can obtain

$$\begin{aligned}V(\bar{x}(k)) &= \bar{x}^T(k)\bar{Q}\bar{x}(k)+\bar{\alpha}u_a^T(k)Ru_a(k)\\ &\quad+\beta\mathbb{E}\left\{\sum_{i=k+1}^{\infty}\beta^{i-k-1}\left(\bar{x}^T(k)\bar{Q}\bar{x}(k)\right.\right.\\ &\quad\left.\left.+\ \alpha(k)u_a^T(k)Ru_a(k)\right)\right\}.\end{aligned} \quad (13)$$

Then, the Bellman equation can be written as

$$\begin{aligned}V(\bar{x}(k)) &= \mathbb{E}\left\{\bar{x}^T(k)\bar{Q}\bar{x}(k)+\alpha(k)u_a^T(k)Ru_a(k)\right.\\ &\quad\left.+\beta V(\bar{x}(k+1))\right\}.\end{aligned} \quad (14)$$

Combining the definition of $V(\bar{x}(k))$, define the following Hamiltonian function

$$\begin{aligned}\mathcal{F}(\bar{x}(k),u_a(k)) &= \bar{x}^T(k)\bar{Q}\bar{x}(k)+\bar{\alpha}u_a^T(k)Ru_a(k)\\ &\quad+\mathbb{E}\left\{\beta\bar{x}^T(k+1)P\bar{x}(k+1)\right\}\\ &\quad-\mathbb{E}\left\{\bar{x}^T(k)P\bar{x}(k)\right\}.\end{aligned} \quad (15)$$

Based on the results in [29], the following equation should be satisfied

$$\begin{aligned}\frac{\partial\mathcal{F}(\bar{x}(k),u_a(k))}{\partial u_a(k)} &= 2\bar{\alpha}Ru_a(k)+2\bar{\alpha}\beta\bar{B}^TP\bar{A}\bar{x}(k)\\ &\quad+2\bar{\alpha}\beta\bar{B}^TP\bar{B}u_a(k)=0,\end{aligned}$$

which implies

$$u_a(k) = -\left(R+\beta\bar{B}^TP\bar{B}\right)^{-1}\beta\bar{B}^TP\bar{A}\bar{x}(k). \quad (16)$$

Submitting (3) and (16) into (14) yields the Riccati equation (12), which completes the proof. ■

If we can obtain the solution P through solving the Riccati equation in (12), the optimal control gain $\bar{K}$ can be designed. However, there exists a variable $\bar{\alpha}$ in the equation (12), which affects the existence and uniqueness of the solution P [38]. It is thus necessary to discuss the relation between $\bar{\alpha}$ and the existence and uniqueness of P.

C. Analysis of the solution to the Riccati equation

The existence of the solution to the Riccati equation (12) will be analyzed in this subsection. Before giving the existence conditions of the solution, define the following functions

$$\begin{aligned}\mathcal{H}(X) &= \bar{Q}+\beta\bar{A}^TX\bar{A}\\ &\quad-\bar{\alpha}\beta^2\bar{A}^TX\bar{B}\left(R+\beta\bar{B}^TX\bar{B}\right)^{-1}\bar{B}^TX\bar{A},\\ \mathcal{H}^k(X) &= \mathcal{H}(\mathcal{H}^{k-1}(X)),\end{aligned} \quad (17)$$

where $\mathcal{H}^k(X)$ means k times composition function for any positive integer k.

Remark 3: Here, (17) is defined to facilitate analyzing the existence and uniqueness of the solution to the Riccati equation in (12). It is obvious that the definition of $\mathcal{H}(X)$ is equivalent to the right side in (12). If we can show that $\mathcal{H}(X)$ can converge to a unique bound, and $H(X)=X$ has a unique solution, the existence and uniqueness of the solution to the Riccati equation (12) can be proved.

The following theorem is proposed to show that $\mathcal{H}(X)$ can converge to a unique bound.

Theorem 2: If there exists a matrix $\tilde{X}\geq 0$ such that the inequality $\mathcal{H}(X)\leq\tilde{X}$ holds, $\mathcal{H}(X)=X$ has a unique solution X_* and $\lim_{k\to\infty}\mathcal{H}^k(X_0)=X_*$ for any initial value $X_0\geq 0$.

Proof: Define the following functions

$$\begin{aligned}\mathcal{H}_1(\bar{K},X) &= \bar{Q}+\beta\left(1-\bar{\alpha}\right)\bar{A}^TX\bar{A}\\ &\quad+\bar{\alpha}\left(\beta(\bar{A}+\bar{B}\bar{K})^TX(\bar{A}+\bar{B}\bar{K})+\bar{K}^TR\bar{K}\right),\end{aligned}$$

$$\mathcal{H}_1^k(\bar{K}, X) \;=\; \mathcal{H}_1(\bar{K}, \mathcal{H}_1^{k-1}(X)). \tag{18}$$

By computing, we can find that $\mathcal{H}_1(\bar{K}, X)$ is equal to $\mathcal{H}(X)$. Accordingly, if we can prove the following conditions hold, the proof can be completed.

If there exists a matrix $\tilde{X} \geq 0$ satisfying $\mathcal{H}_1(\bar{K}, \tilde{X}) \leq \tilde{X}$, $\mathcal{H}_1(\bar{K}, X) = X$ has a unique solution X_* and $\lim_{k\to\infty} \mathcal{H}_1^k(\bar{K}, X_0) = X_*$ for any initial value $X_0 \geq 0$.

Firstly, as can be seen from the structure of $\mathcal{H}_1(\bar{K}, X)$, it is a monotonically increasing function w.r.t the variable X. For the zero initial condition $X_0 = 0$, $X_k = \mathcal{H}_1^k(\bar{K}, X_0)$. According to the monotonically increasing property of the function $\mathcal{H}_1^k(\cdot)$, the following monotonic sequence can be obtained

$$0 \;=\; X_0 < X_1 \leq \ldots \leq X_k.$$

Similarly, iteratively using the condition $\mathcal{H}_1(\bar{K}, \tilde{X}) \leq \tilde{X}$ yields

$$\mathcal{H}_1^k(\bar{K}, \tilde{X}) \;\leq\; \ldots \leq \mathcal{H}_1(\bar{K}, \tilde{X}) \leq \tilde{X}.$$

Considering the fact $\tilde{X} > X_0$ and the monotonically increasing property of the function $\mathcal{H}_1^k(\cdot)$, we can obtain

$$X_k \;\leq\; \mathcal{H}_1^k(\bar{K}, \tilde{X}) \leq \tilde{X},$$

which implies the sequence X_k is monotonically increasing and bounded. In this way, we conclude that the sequence X_k can converge to X_*.

Next, consider a general case, that is, the initial condition is for $X_0 \geq 0$. Then, we can always find a positive scalar η such that $X_* \geq \eta X_0$ holds. Combining the above results, the following inequality can be derived

$$\mathcal{H}_1^k(\bar{K}, 0) \quad\leq\quad \mathcal{H}_1^k(\bar{K}, \eta X_0) \leq \mathcal{H}_1^k(\bar{K}, X_*) \leq X_*.$$

Therefore, $\lim_{k\to\infty} \mathcal{H}_1^k(\bar{K}, \eta X_0) = X_*$ holds. According to the structure, the following equation holds

$$\mathcal{H}_1^k(\bar{K}, \eta X_0) - \mathcal{H}_1^k(\bar{K}, 0) = \eta \left(\mathcal{H}_1^k(\bar{K}, X_0) - \mathcal{H}_1^k(\bar{K}, 0)\right),$$

which implies

$$\lim_{k\to\infty} \mathcal{H}_1^k(\bar{K}, X_0) \quad=\quad X_*.$$

Using the similar approach, we can conclude that $\mathcal{H}(X) = X$ has a unique solution X_* and $\lim_{k\to\infty} \mathcal{H}^k(X_0) = X_*$ for any initial value $X_0 \geq 0$. The proof is completed. ■

Different from the general Riccati equation, the one in (12) has a parameter $\bar{\alpha}$, which will affect its convergence. Namely, if the adversary can implement the attacks with a high probability, the conditions in Theorem 2 cannot be satisfied anymore nor does the solution to the Riccati equation exist. Therefore, Theorem 3 is proposed to show that on what conditions the solution to the Riccati equation exists.

Theorem 3: For the case $|\sqrt{\beta}\rho| > 1$, the following inequality is necessary to ensure that the solution to the Riccati equation (12) exists

$$\beta\,(1-\bar{\alpha}) \leq \frac{1}{\rho^2},$$

where ρ means the spectral radius of the matrix $\bar{A}$.

Proof: According to Theorem 2, we know that there exists a matrix $X \geq 0$ satisfying

$$\begin{aligned} X \;\geq\; & \mathcal{H}(X) = \bar{Q} + \beta \bar{A}^T X \bar{A} \\ & -\bar{\alpha}\beta^2 \bar{A}^T X \bar{B} \left(R + \beta \bar{B}^T X \bar{B}\right)^{-1} \bar{B}^T X \bar{A} \\ =\; & \bar{Q} + \beta\,(1-\bar{\alpha})\,\bar{A}^T X \bar{A} + \bar{\alpha}\beta \bar{A}^T X \bar{A} \\ & -\bar{\alpha}\beta^2 \bar{A}^T X \bar{B} \left(R + \beta \bar{B}^T X \bar{B}\right)^{-1} \bar{B}^T X \bar{A}. \end{aligned} \tag{19}$$

Based on the matrix inverse lemma [39], the following equation holds

$$\begin{aligned} & \bar{\alpha}\beta \bar{A}^T X \bar{A} - \bar{\alpha}\beta^2 \bar{A}^T X \bar{B} \left(R + \beta \bar{B}^T X \bar{B}\right)^{-1} \bar{B}^T X \bar{A} \\ & = \bar{\alpha}\beta \bar{A}^T \left(X^{-1} + \beta \bar{B} R^{-1} \bar{B}^T\right)^{-1} \bar{A}. \end{aligned} \tag{20}$$

Submitting (20) in (19) yields

$$\begin{aligned} X \;\geq\; & \bar{Q} + \beta\,(1-\bar{\alpha})\,\bar{A}^T X \bar{A} \\ & + \bar{\alpha}\beta \bar{A}^T \left(X^{-1} + \beta \bar{B} R^{-1} \bar{B}^T\right)^{-1} \bar{A}. \end{aligned}$$

Since $\bar{B}$ is not invertible and $X > 0$, $\bar{\alpha}\beta \bar{A}^T (X^{-1} + \beta \bar{B} R^{-1} \bar{B}^T)^{-1} \bar{A} \geq 0$. Thus,

$$X \;\geq\; \bar{Q} + \beta\,(1-\bar{\alpha})\,\bar{A}^T X \bar{A},$$

which shows that $\beta\,(1-\bar{\alpha}) \leq \frac{1}{\rho^2}$ must hold. The proof is completed. ■

Remark 4: It is worth noting that the condition in Theorem 3 depends on ρ. We, however, do not know the exact matrix $\bar{A}$. In the next section, we will design the controller via using the Q-learning approach, based on which we can obtain ρ.

Based on the above results, the following theorem is proposed to show the stability of the augmented system (3) and the optimality of the controller (11) is guaranteed by the following theorem, whose proof is omitted for want of space; see [29].

Theorem 4: For the augmented system (3), Theorems 1-3 hold. Define $\bar{e}(k) = \beta^{\frac{k}{2}} e(k)$ with $e(k) = y(k) - r(k)$. The optimal tracking controller in (11) can stabilize $\bar{e}(k)$. Meanwhile, the value function $V(k) = \mathbb{E}\left\{\bar{x}^T(k) P \bar{x}(k)\right\}$ can be minimized.

4.. Q-LEARNING OPTIMAL TRACKING SCHEME DESIGN

The above results show that the optimal tracking scheme exists yet the exact system dynamics are needed in the design process. Accordingly, the Q-learning approach is introduced to design a model-free optimal tracking control scheme. To facilitate analyzing the convergence of the tracking control algorithm without using system knowledge, an algorithm with the system parameters is designed. Based on such an algorithm, the Q-learning tracking control algorithm without using system knowledge is provided.

Define the Q-function as follows

$$\begin{aligned} Q(\bar{x}(k), u_a(k)) \;=\; & \mathbb{E}\left\{\bar{x}^T(k) \bar{Q} \bar{x}(k) + \alpha(k) u_a^T(k) R u_a(k) \right. \\ & \left. + \beta \bar{x}^T(k+1) P \bar{x}(k+1)\right\}. \end{aligned} \tag{21}$$

Using (3), (21) can be rewritten as

$$Q(\xi(k)) \;=\; \xi^T(k) \mathcal{M} \xi(k) + \mathbb{E}\left\{\xi^T(k) \begin{bmatrix} \bar{A}^T \\ \alpha(k) \bar{B}^T \end{bmatrix}\right.$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3056633, IEEE Internet of Things Journal

$$
\begin{aligned}
&\times P \left[\begin{array}{c} \bar{A}^T \\ \alpha(k)\bar{B}^T \end{array} \right]^T \xi(k) \Bigg\} \\
&= \xi^T(k) \left[\begin{array}{cc} \bar{Q}+\beta\bar{A}^T P\bar{A} & \bar{\alpha}\beta\bar{A}^T P\bar{B} \\ * & \bar{\alpha}R+\bar{\alpha}\beta\bar{B}^T P\bar{B} \end{array} \right] \xi(k) \\
&= \xi^T(k)\mathcal{Q}\xi(k), \qquad (22)
\end{aligned}
$$

where

$$
\begin{aligned}
\xi(k) &= \left[\begin{array}{c} \bar{x}(k) \\ u_a(k) \end{array} \right], \ \mathcal{Q} = \left[\begin{array}{cc} \mathcal{Q}_{11} & \mathcal{Q}_{12} \\ * & \mathcal{Q}_{22} \end{array} \right], \\
\mathcal{Q}_{11} &= \bar{Q}+\beta\bar{A}^T P\bar{A}, \ \mathcal{Q}_{12} = \bar{\alpha}\beta\bar{A}^T P\bar{B}, \\
\mathcal{Q}_{22} &= \bar{\alpha}R+\bar{\alpha}\beta\bar{B}^T P\bar{B}, \ \mathcal{M} = \text{diag}\left\{\bar{Q}, \bar{\alpha}R\right\}.
\end{aligned}
$$

Based on [25], we know that the relation between P and $\mathcal{Q}$ can be described as

$$
P = \left[\begin{array}{cc} I & \bar{K}^T \end{array} \right] \mathcal{Q} \left[\begin{array}{cc} I & \bar{K}^T \end{array} \right]^T,
$$

which further implies

$$
\begin{aligned}
\mathcal{Q} &= \mathcal{M} + \mathbb{E}\left\{ \left[\begin{array}{cc} \bar{A}^T & \bar{A}^T\bar{K}^T \\ \alpha(k)\bar{B}^T & \alpha(k)\bar{B}^T\bar{K}^T \end{array} \right] \right. \\
&\left. \times \mathcal{Q} \left[\begin{array}{cc} \bar{A} & \alpha(k)\bar{B} \\ \bar{K}\bar{A} & \alpha(k)\bar{K}\bar{B} \end{array} \right] \right\}. \qquad (23)
\end{aligned}
$$

Based on (16), the optimal tracking controller can be thus described as

$$
u_a(k) = -\mathcal{Q}_{22}^{-1}\mathcal{Q}_{12}^T\bar{x}(k). \qquad (24)
$$

Accordingly, if we can learn online and obtain the matrix $\mathcal{Q}$, the optimal controller can be designed without using the system dynamics.

Based on (21) and (22), the Q-function satisfies the following equation

$$
\begin{aligned}
\xi^T(k)\mathcal{Q}\xi(k) &= \xi^T(k)\mathcal{M}\xi + \mathbb{E}\left\{\beta Q(\bar{x}(k+1), u_a(k+1))\right\} \\
&= \xi^T(k)\mathcal{M}\xi \\
&\quad + \mathbb{E}\left\{\beta\xi^T(k+1)\mathcal{Q}\xi(k+1)\right\}. \qquad (25)
\end{aligned}
$$

Define $\bar{\xi}(k) = \xi^T(k) \otimes \xi^T(k)$. (25) is equivalent to

$$
\bar{\xi}(k)\text{vec}(\mathcal{Q}) = \bar{\xi}(k)\text{vec}(\mathcal{M}) + \beta\mathbb{E}\left\{\bar{\xi}(k+1)\text{vec}(\mathcal{Q})\right\} \qquad (26)
$$

The least-squares approach can be used to obtain $\mathcal{Q}$. Note that $\mathcal{Q}$ is a $(n_x+n_u+n_y)\times(n_x+n_u+n_y)$ symmetric matrix. Therefore, at least $n = (n_x+n_u+n_y)(n_x+n_u+n_y+1)/2$ data should be provided to solve (26).

Now, we are in the position to propose Algorithm 1, which online learns the matrix $\mathcal{Q}$ and obtains the optimal tracking controller.

In the process of fulfilling Algorithm 1, the probing noise $e(k)$ should be added to the control signal in (24), that is, $u_a(k) = -\mathcal{Q}_{22}^{-1}\mathcal{Q}_{12}^T\bar{x}(k) + e(k)$ is actually applied to (25) to generate data, with which the existence of $\left(\bar{\xi}^T\bar{\xi}\right)^{-1}$ can be guaranteed [25].

Remark 5: By implementing Algorithm 1, not only the tracking controller $u_a(k)$ but also the matrix $\mathcal{Q}$ can be obtained. The structure of $\mathcal{Q}$ in (22) implies $\mathcal{Q}_{11} = \bar{Q} + \beta\bar{A}^T P\bar{A} > 0$, which further implies $\mathcal{Q}_{11} - \bar{Q}$ is a positive-definite/semi positive-definite and symmetric matrix. Then, $\mathcal{Q}_{11} - \bar{Q}$ can be decomposed as $\mathbb{D}^T\mathbb{D}$ with $\mathbb{D}$ being a full rank matrix. Also, the matrix P can be decomposed as $\mathbb{M}^T\mathbb{M}$ with $\mathbb{M}$ being a full rank matrix. Then, $\mathcal{Q}_{11} = \bar{Q} + \beta\bar{A}^T P\bar{A}$ is described as

$$
\mathbb{D}^T\mathbb{D} = \beta\bar{A}^T\mathbb{M}^T\mathbb{M}\bar{A},
$$

which implies

$$
\bar{A} = \frac{1}{\beta}\mathbb{M}^{-1}\mathbb{D}.
$$

Moreover, ρ in Theorem 3 is obtained without knowing the exact system dynamics.

Algorithm 1

1: Provide initial values for $\bar{K}_0$ and $\mathcal{Q}_0$
2: Set the allowed learning error ϵ and $i = 0$
3: Execute the policy evaluation
$\xi^T(k)\mathcal{Q}_{i+1}\xi(k) = \bar{x}^T(k)\bar{Q}\bar{x}(k) + \bar{\alpha}u_a^T(k)Ru_a(k) + \mathbb{E}\left\{\beta\xi^T(k+1)\mathcal{Q}_i\xi(k+1)\right\}$
4: Execute the policy improvement
$u_a(k) = -\mathcal{Q}_{22,i+1}^{-1}\mathcal{Q}_{12,i+1}^T\bar{x}(k) + e(k)$
5: Construct $\text{vec}(\mathcal{Q}_{i+1})$
6: **if** $\|\mathcal{Q}_{i+1} - \mathcal{Q}_i\| < \epsilon$ **then**
7: Output the matrix $\mathcal{Q}_{i+1}$ and the optimal tracking controller $u_a(k)$
8: **else**
9: $i = i + 1$
10: Reture to Step 3
11: **end if**

A. Convergence analysis of Algorithm

To prove the convergence of Algorithm 1, the following lemma is given.

Lemma 1: [25] Iterating $\mathcal{Q}_i$ is equivalent to iterating P_i.

Using the conclusion in Lemma 1, the following theorem is proposed.

Theorem 5: If the Riccati equation (12) is solvable, $\mathcal{Q}_i$ in Algorithm 1 can converge to the value $\mathcal{Q}$ with an allowed error ϵ and the optimal tacking controller gain $\bar{K}$ can be obtained.

Proof: Lemma 1 implies whether $\mathcal{Q}_i$ converges or not depends on the convergence of P_i. Theorem 2 proves that the Riccati equation (12) is solvable. Thus $\mathcal{Q}_i$ can converge and the gain $\bar{K}$ can be obtained. The proof is completed. ■

Next, a dc motor is adopted to validate the proposed optimal tracking control scheme. Both the model-based and model-free simulation results are provided. In this way, the effectiveness of the proposed scheme can be clearly shown.

Example 1: The system matrices for the dc motor are given as follows [40]

$$
\begin{aligned}
A &= \left[\begin{array}{cc} 1.00021 & 0.00460 \\ 0.00460 & 0.00004 \end{array} \right], \ B = \left[\begin{array}{c} 0.34868 \\ 7.68069 \end{array} \right], \\
C &= \left[\begin{array}{cc} 1 & 0 \end{array} \right].
\end{aligned}
$$

For the dc motor, regard the voltage and the angular position as the control input and output, respectively. First, we use Theorems 1 and 2 to calculate the solution P and the optimal gain $\bar{K}$. Set $Q = 2$, $R = 2$, the discounted factor $\beta = 0.7$,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3056633, IEEE Internet of Things Journal

the reference model gain $G = 1$. Then, the solution P and the optimal gain K are computed as

$$P = \begin{bmatrix} 5.2114 & 0.0148 & -5.2096 \\ 0.0148 & 0.0001 & -0.0148 \\ -5.2096 & -0.0148 & 5.2077 \end{bmatrix},$$
$$\bar{K} = \begin{bmatrix} -0.5403 & -0.0025 & 0.5400 \end{bmatrix}.$$

The reference trajectory is set as $r(k) = 3$ rad. Figs. 2 and 3 show the simulation results of the proposed tracking control scheme. According to these two figures, we can conclude the proposed tracking controller can track the desired trajectory under attacks. However, DoS attacks can deteriorate the tracking performance.

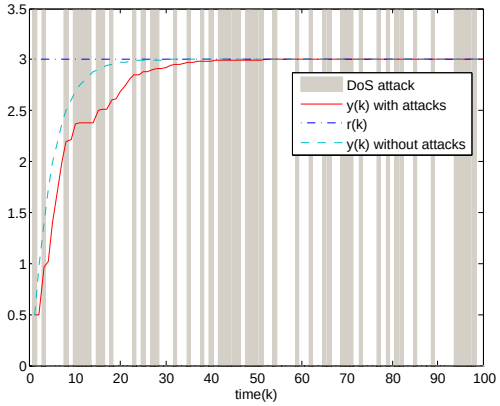

Fig. 2. Tracking performance comparisons under attacks and without attacks.

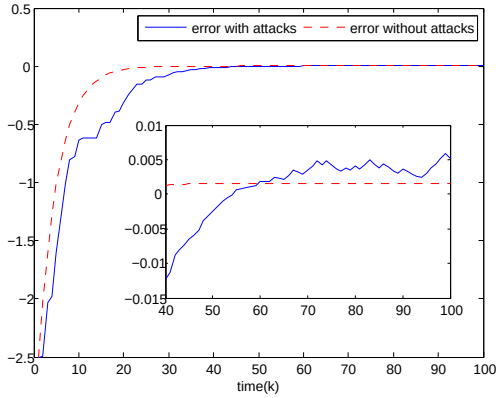

Fig. 3. Tracking errors under attacks and without attacks.

Next, the simulation results of the model-free optimal tracking control scheme are given. According to (22), we first compute the matrix $\mathcal{Q}$ to be learned as

$$\mathcal{Q} = \begin{bmatrix} 5.6496 & 0.0168 & -5.6475 & 0.8110 \\ 0.0168 & 0.0001 & -0.0168 & 0.0037 \\ -5.6475 & -0.0168 & 5.6454 & -0.8105 \\ 0.8110 & 0.0037 & -0.8105 & 1.5010 \end{bmatrix}.$$

To apply Algorithm 1 to design the tracking control scheme, 30 data is collected for each iteration and a probing noise is added to fully explore the state space. Figs. 4 and 5 depict the convergence of the matrix $\mathcal{Q}_i$ and the control gain $\bar{K}_i$, respectively. As can be seen from Figs. 4 and 5, $\mathcal{Q}_i$ and $\bar{K}_i$ can converge to the desired values after 20 iterations. The values are respectively as

$$\mathcal{Q}_{20} = \begin{bmatrix} 5.6496 & 0.0168 & -5.6475 & 0.8110 \\ 0.0168 & 0.0001 & -0.0168 & 0.0037 \\ -5.6475 & -0.0168 & 5.6454 & -0.8105 \\ 0.8110 & 0.0037 & -0.8105 & 1.5010 \end{bmatrix},$$
$$\bar{K}_{20} = \begin{bmatrix} -0.5403 & -0.0025 & 0.5400 \end{bmatrix}.$$

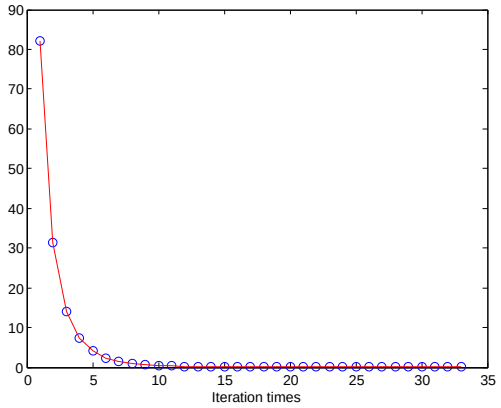

Fig. 4. The error of $\|\mathcal{Q}_i - \mathcal{Q}\|$.

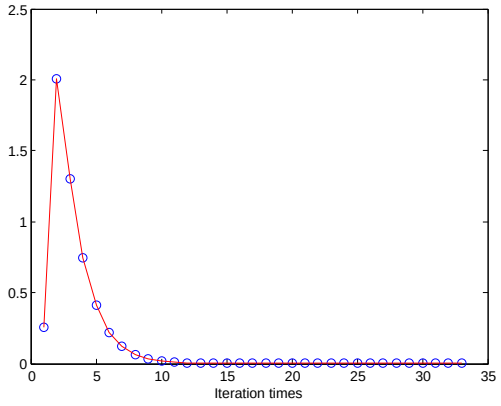

Fig. 5. The error of $\|\bar{K}_i - \bar{K}\|$.

Using Algorithm 1, the optimal tracking control scheme is implemented. Fig. 6 shows the responses of the output signal and the reference signal. The added probing noise is depicted in Fig. 7. It is clear that the probing noise is not added any more after the control gain is successfully learned online and the output signal can be driven to track the given reference signal.

To show that the DoS attack can increase the control cost and affect the learning rate, Tab. I gives different values for $\|P\|$. It can be seen that $\|P\|$ increases as $\bar{\alpha}$ decreases. Fig. 8 demonstrates that along with the improvement of cyber-layer security, the learning needs fewer and fewer iteration times.

Although the approach proposed in Algorithm 1 can design the optimal tracking controller without using the exact system dynamics, a bias of the solution can happen due to adding the probing noise $e(k)$ in the control signal, which may result in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3056633, IEEE Internet of Things Journal

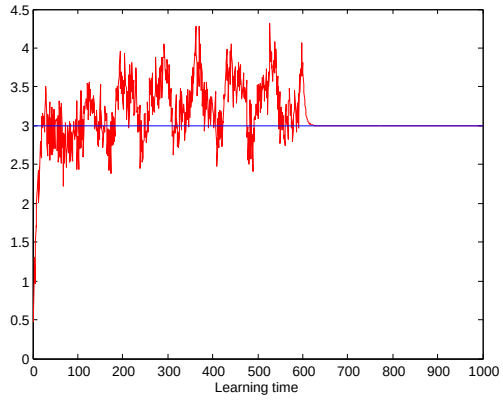

Fig. 6. The trajectories of the output $y(k)$ and reference signal $r(k)$ in the learning.

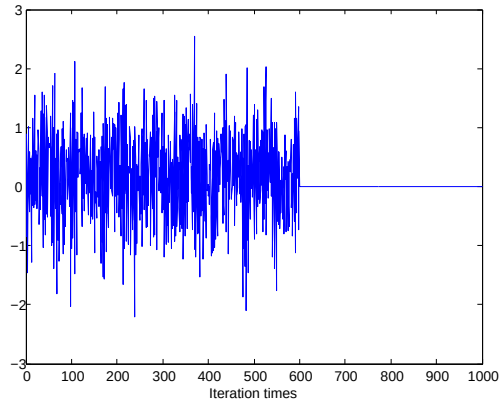

Fig. 7. The response of the added probing noise in the learning.

TABLE I
$\|P\|$ UNDER DIFFERENT $\bar{\alpha}$.

$\bar{\alpha}$	0.6	0.7	0.8	0.9	1
$\|P\|$	10.4192	10.1152	9.8398	9.5886	9.3583

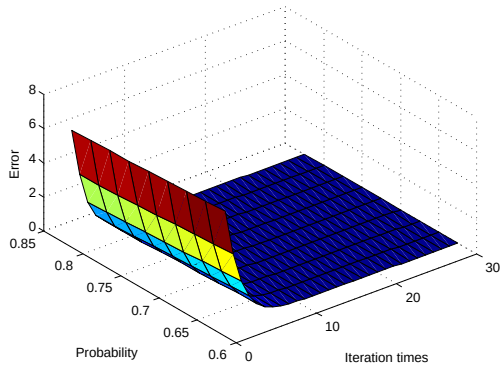

Fig. 8. The learning times under different probability $\bar{\alpha}$. ("Error" denotes $\|\mathcal{Q}_i - \mathcal{Q}\|$.)

an incorrect control gain $\bar{K}$. In the following, we show the bias resulting from adding the probing noise in Algorithm 1.

When adding the probing noise $e(k)$, the following equation can be obtained

$$\begin{aligned}\xi^T(k)\mathcal{Q}_{i+1}\xi(k) &= \bar{x}^T(k)\bar{Q}\bar{x}(k)+\bar{\alpha}u_a^T(k)Ru_a(k)\\ &\quad+\mathbb{E}\left\{\beta\xi^T(k+1)\mathcal{Q}_i\xi(k+1)\right\}\\ &= \bar{x}^T(k)\bar{Q}\bar{x}(k)+\bar{\alpha}u_a^T(k)Ru_a(k)\\ &\quad+2\bar{\alpha}e^T(k)Re(k)+2\bar{\alpha}u_a^T(k)Re(k)\\ &\quad+\mathbb{E}\left\{\beta\xi^T(k+1)\mathcal{Q}_i\xi(k+1)\right\}.\end{aligned}$$

Compared with Line 3 in Algorithm 1, two extra items depending on $e(k)$, that is, $2\bar{\alpha}e^T(k)Re(k)$ and $2\bar{\alpha}u_a^T(k)Re(k)$ exist in the above equation, which can make the solution incorrect. To avoid such a problem, we will provide an alternative approach to designing the tracking control scheme without using the exact system dynamics. The policy $u_a(k)$ to be updated is also applied to the system to generate data for learning in Algorithm 1, which is viewed as an on-policy approach [21]. In the following, an off-policy learning approach is presented to realize the control objective of this paper.

5.. OFF-POLICY LEARNING CONTROL ALGORITHM

This section mainly investigates how to propose an off-policy learning control scheme for CPS under DoS attacks without knowing the complete system dynamics. The physical process in (1) under actuator DoS attacks is rewritten as

$$x(k+1) = \bar{A}_i x(k)+\alpha(k)\bar{B}\left(\bar{K}_i x(k)+u_a(k)\right). \quad (27)$$

where $\bar{A}_i = \bar{A}-\alpha(k)\bar{B}\bar{K}_i$. $u_a^i(k) = -\bar{K}_i x(k)$ is the target policy to be learned and updated and $i\in\mathbb{Z}$ is each learning step.

For the learned $u_a^i(k)$, the Bellman equation (14) can be described as

$$\begin{aligned}&V^{i+1}(\bar{x}(k),u_a(k))-\mathbb{E}\left\{\beta V^{i+1}(\bar{x}(k+1),u_a(k))\right\}\\ =\ &\mathbb{E}\left\{\bar{x}^T(k)\bar{Q}\bar{x}(k)+\alpha(k)u_a^T(k)Ru_a(k)\right\}. \quad (28)\end{aligned}$$

At the point $\bar{x}(k+1)$, the Taylor expansion of $V(\bar{x}(k)$ can be calculated as

$$\begin{aligned}V(\bar{x}(k) &= \mathbb{E}\left\{V(\bar{x}(k+1)+2\bar{x}^T(k+1)P\right.\\ &\quad\times(\bar{x}(k)-\bar{x}(k+1))+(\bar{x}(k)-\bar{x}(k+1))^T\\ &\quad\left.\times P\left(\bar{x}(k)-\bar{x}(k+1)\right)\right\}. \quad (29)\end{aligned}$$

Then, (28) can be rewritten as

$$\begin{aligned}&V^{i+1}(\bar{x}(k),u_a(k))-\mathbb{E}\left\{\beta V^{i+1}(\bar{x}(k+1),u_a(k))\right\}\\ =\ &\mathbb{E}\left\{\bar{x}^T(k)P_{i+1}\bar{x}(k)-\beta\bar{x}^T(k)\bar{A}_i^TP_{i+1}\bar{A}_i\bar{x}(k)\right.\\ &-\alpha(k)\beta\left(\bar{K}_ix(k)+u_a(k)\right)^T\bar{B}^TP_{i+1}\bar{x}(k+1)\\ &\left.-\alpha(k)\beta\left(\bar{K}_ix(k)+u_a(k)\right)^T\bar{B}^TP_{i+1}\bar{A}_i\bar{x}(k)\right\}. \quad (30)\end{aligned}$$

According to (27) and the Bellman equation (14), the following Lyapunov equation can be obtained

$$\bar{Q}-P_{i+1}+\bar{\alpha}\bar{K}_i^TR\bar{K}_i^T+\mathbb{E}\left\{\beta\bar{A}_i^TP_{i+1}\bar{A}_i\right\} = 0. \quad (31)$$

Combining $V(\bar{x}(k)) = \mathbb{E}\left\{\bar{x}^T(k)P\bar{x}(k)\right\}$, (30) and (31) yields

$$\begin{aligned}&\mathbb{E}\left\{\bar{x}^T(k)P_{i+1}\bar{x}(k) - \beta\bar{x}^T(k+1)P_{i+1}\bar{x}(k+1)\right\}\\ =\ & \bar{x}^T(k)\bar{Q}\bar{x}(k) + \bar{\alpha}\bar{x}^T(k)\bar{K}_i^T R\bar{K}_i^T\bar{x}(k)\\ &-\mathbb{E}\left\{\alpha(k)\beta\left(\bar{K}_i x(k) + u_a(k)\right)^T\bar{B}^T P_{i+1}\bar{x}(k+1)\right\}\\ &-\mathbb{E}\left\{\alpha(k)\beta\left(\bar{K}_i x(k) + u_a(k)\right)^T\bar{B}^T P_{i+1}\bar{A}_i\bar{x}(k)\right\}\end{aligned} \tag{32}$$

A. Model-based off-policy learning control scheme

If we provide a stabilizing control signal $u_a(k)$ and an initial $\bar{K}_0$ for (32), the solution P_{i+1} and $\bar{K}_{i+1}$ can be solved iteratively using the least-square approach. The details are described in Algorithm 2.

Algorithm 2

1: Set the initial learning step $i = 0$ and the learning error ϵ
2: Give an admissible controller $u_a(k)$
3: Obtain $\bar{K}_{i+1}$, P_{i+1} through solving (32) using the least-square approach
4: **if** $\|\bar{K}_{i+1} - \bar{K}_i\| < \epsilon$ **then**
5: Output $\bar{K}_{i+1}$ as the optimal control gain
6: **else**
7: $i = i + 1$
8: Reture to Step 3
9: **end if**

Next, the following theorem is proposed to show the convergence of Algorithm 2.

Theorem 6: The gain $\bar{K}_{i+1}$ obtained from Algorithm 2 can converge to the optimal control scheme (11).

Proof: Submitting $\bar{A}_i$ and (1) into (32) yields

$$\begin{aligned}&\mathbb{E}\left\{\bar{x}^T(k)P_{i+1}\bar{x}(k) - \beta\left(\bar{A}\bar{x}(k) + \alpha(k)\bar{B}u_a(k)\right)^T\right.\\ &\left.\times P_{i+1}\left(\bar{A}\bar{x}(k) + \alpha(k)\bar{B}u_a(k)\right)\right\}\\ =\ & \bar{x}^T(k)\bar{Q}\bar{x}(k) + \bar{\alpha}\bar{x}^T(k)\bar{K}_i^T R\bar{K}_i^T\bar{x}(k)\\ &-\mathbb{E}\left\{\alpha(k)\beta\left(\bar{K}_i x(k) + u_a(k)\right)^T\bar{B}^T P_{i+1}\right.\\ &\left.\times\left(\bar{A}\bar{x}(k) + \alpha(k)\bar{B}u_a(k)\right)\right\}\\ &-\mathbb{E}\left\{\alpha(k)\beta\left(\bar{K}_i x(k) + u_a(k)\right)^T\right.\\ &\left.\times\bar{B}^T P_{i+1}\left(\bar{A} - \alpha(k)\bar{B}\bar{K}_i\right)\bar{x}(k)\right\}.\end{aligned} \tag{33}$$

By direct calculation, (33) can be obtained as

$$\begin{aligned}&\bar{x}^T(k)P_{i+1}\bar{x}(k) - \beta\bar{x}^T(k)\bar{A}^T P_{i+1}\bar{A}\bar{x}(k)\\ =\ & \bar{x}^T(k)\bar{Q}\bar{x}(k) + \bar{\alpha}\bar{x}^T(k)\bar{K}_i^T R\bar{K}_i^T\bar{x}(k)\\ &-2\bar{\alpha}\beta\bar{x}^T(k)\bar{K}_i^T\bar{B}P_{i+1}\bar{A}\bar{x}(k)\\ &+\bar{\alpha}\beta\bar{x}^T(k)\bar{K}_i^T\bar{B}P_{i+1}\bar{B}\bar{K}_i\bar{x}(k),\end{aligned}$$

which further implies

$$\begin{aligned}P_{i+1} =\ & \bar{Q} + \beta\bar{A}^T P_{i+1}\bar{A} - \bar{\alpha}\beta^2\bar{A}^T P_{i+1}\\ &\times\bar{B}\left(R + \beta\bar{B}^T P_{i+1}\bar{B}\right)^{-1}\bar{B}^T P_{i+1}\bar{A}.\end{aligned} \tag{34}$$

Theorem 2 concludes that P_i can converge. Thus the gain $\bar{K}_i$ solved using Algorithm 2 can converge to the desired optimal value. The proof is completed. ■

Before analyzing the effect of the probing noise $e(k)$ on the solution obtained from Algorithm 2, define $\tilde{P}_{i+1}$ and $\hat{P}_{i+1}$ respectively as the solutions under $e(k) \neq 0$ and $e(k) = 0$. Next, the following theorem shows that $\tilde{P}_{i+1} = \hat{P}_{i+1}$.

Theorem 7: The solution $\tilde{P}_{i+1}$ under the probing noise is equal to $\hat{P}_{i+1}$ without the probing noise.

Proof: Under $e(k) \neq 0$, (32) can be described as

$$\begin{aligned}&\mathbb{E}\left\{\bar{x}^T(k)\tilde{P}_{i+1}\bar{x}(k) - \beta\left(\bar{x}(k+1) + \alpha(k)\bar{B}e(k)\right)^T\right.\\ &\left.\times\tilde{P}_{i+1}\left(\bar{x}(k+1) + \alpha(k)\bar{B}e(k)\right)\right\}\\ =\ & \bar{x}^T(k)\bar{Q}\bar{x}(k) + \bar{\alpha}\bar{x}^T(k)\bar{K}_i^T R\bar{K}_i^T\bar{x}(k)\\ &-\mathbb{E}\left\{\alpha(k)\beta\left(\bar{K}_i x(k) + u_a(k) + e(k)\right)^T\bar{B}^T\tilde{P}_{i+1}\right.\\ &\left.\times\left(\bar{A}_i\bar{x}(k) + \alpha(k)\bar{B}\left(u_a(k) + e(k)\right)\right)\right\}\\ &-\mathbb{E}\left\{\alpha(k)\beta\left(\bar{K}_i x(k) + u_a(k) + e(k)\right)^T\right.\\ &\left.\times\bar{B}^T\tilde{P}_{i+1}\left(\bar{A} - \alpha(k)\bar{B}\bar{K}_i\right)\bar{x}(k)\right\}.\end{aligned} \tag{35}$$

By performing some mathematical operations, (35) can be written as

$$\begin{aligned}&\bar{x}^T(k)\tilde{P}_{i+1}\bar{x}(k) - \mathbb{E}\left\{\beta\bar{x}^T(k+1)\tilde{P}_{i+1}\bar{x}(k+1)\right\}\\ &-2\mathbb{E}\left\{\alpha(k)\beta\bar{x}^T(k+1)\tilde{P}_{i+1}\bar{B}e(k)\right\}\\ &-\bar{\alpha}\beta e^T(k)\bar{B}^T\tilde{P}_{i+1}\bar{B}e(k)\\ =\ & \bar{x}^T(k)\bar{Q}\bar{x}(k) + \bar{\alpha}\bar{x}^T(k)\bar{K}_i^T R\bar{K}_i^T\bar{x}(k)\\ &-\mathbb{E}\left\{\alpha(k)\beta\left(\bar{K}_i x(k) + u_a(k)\right)^T\bar{B}^T\tilde{P}_{i+1}\bar{x}(k+1)\right\}\\ &-\mathbb{E}\left\{\alpha(k)\beta\left(\bar{K}_i x(k) + u_a(k)\right)^T\bar{B}^T\tilde{P}_{i+1}e(k)\right\}\\ &-\mathbb{E}\left\{\alpha(k)\beta\bar{x}^T(k+1)\tilde{P}_{i+1}\bar{B}e(k)\right\}\\ &-\bar{\alpha}\beta e^T(k)\bar{B}^T\tilde{P}_{i+1}\bar{B}e(k)\\ &-\mathbb{E}\left\{\alpha(k)\beta\left(\bar{K}_i x(k) + u_a(k)\right)^T\bar{B}^T\tilde{P}_{i+1}\bar{A}_i\bar{x}(k)\right\}\\ &-\bar{\alpha}\beta e^T(k)\bar{B}^T\tilde{P}_{i+1}\bar{A}_i\bar{x}(k),\end{aligned}$$

which implies

$$\begin{aligned}&\bar{x}^T(k)\tilde{P}_{i+1}\bar{x}(k) - \mathbb{E}\left\{\beta\bar{x}^T(k+1)\tilde{P}_{i+1}\bar{x}(k+1)\right\}\\ =\ & \bar{x}^T(k)\bar{Q}\bar{x}(k) + \bar{\alpha}\bar{x}^T(k)\bar{K}_i^T R\bar{K}_i^T\bar{x}(k)\\ &-\mathbb{E}\left\{\alpha(k)\beta\left(\bar{K}_i x(k) + u_a(k)\right)^T\bar{B}^T\tilde{P}_{i+1}\bar{x}(k+1)\right\}\\ &-\mathbb{E}\left\{\alpha(k)\beta\left(\bar{K}_i x(k) + u_a(k)\right)^T\bar{B}^T\tilde{P}_{i+1}\bar{A}_i\bar{x}(k)\right\}\end{aligned} \tag{36}$$

Thus, $\tilde{P}_{i+1}$ can be solved from (36).

When $e(k) = 0$, (32) is rewritten as

$$\begin{aligned}&\bar{x}^T(k)\hat{P}_{i+1}\bar{x}(k) - \mathbb{E}\left\{\beta\bar{x}^T(k+1)\hat{P}_{i+1}\bar{x}(k+1)\right\}\\ =\ & \bar{x}^T(k)\bar{Q}\bar{x}(k) + \bar{\alpha}\bar{x}^T(k)\bar{K}_i^T R\bar{K}_i^T\bar{x}(k)\\ &-\mathbb{E}\left\{\alpha(k)\beta\left(\bar{K}_i x(k) + u_a(k)\right)^T\bar{B}^T\hat{P}_{i+1}\bar{x}(k+1)\right\}\\ &-\mathbb{E}\left\{\alpha(k)\beta\left(\bar{K}_i x(k) + u_a(k)\right)^T\bar{B}^T\hat{P}_{i+1}\bar{A}_i\bar{x}(k)\right\}\end{aligned} \tag{37}$$

$\hat{P}_{i+1}$ can be solved from (37). As can be seen from (36) and (37), $\tilde{P}_{i+1} = \hat{P}_{i+1}$ holds, which shows that adding the probing noise in Algorithm 2 will not result in the bias of the solution.

Thus, the gain $\bar{K}_{i+1}$ obtained from Algorithm 2 can converge to the optimal control scheme (11). The proof is completed. ■

B. Model-free off-policy learning control scheme

It is easy to see that system dynamics are necessary information in Algorithm 2. Next, the off-policy learning control scheme without using exact system dynamics is proposed based on Algorithm 2.

Based on Kronecker product, (32) can be described as

$$\begin{aligned}
&\left(\bar{x}^T(k)\otimes\bar{x}^T(k)\right)\mathrm{vec}(P_{i+1})\\
&-\beta\mathbb{E}\left\{\left(\bar{x}^T(k+1)\otimes\bar{x}^T(k+1)\right)\mathrm{vec}(P_{i+1})\right\}\\
&+2\bar{\alpha}\beta\left(\bar{x}^T(k)\otimes\left(\bar{K}_i\bar{x}(k)+u_a(k)\right)^T\right)\mathrm{vec}(\bar{B}^TP_{i+1}\bar{A})\\
&-\bar{\alpha}\beta\left(\left(\bar{K}_i\bar{x}(k)-u_a(k)\right)^T\otimes\left(\bar{K}_i\bar{x}(k)+u_a(k)\right)^T\right)\\
&\times\mathrm{vec}(\bar{B}^TP_{i+1}\bar{B})\\
=\;&\bar{x}^T(k)\bar{Q}\bar{x}(k)+\bar{\alpha}\bar{x}^T(k)\bar{K}_i^TR\bar{K}_i^T\bar{x}(k). \qquad (38)
\end{aligned}$$

By direct mathematical operations, we can see that the following equation is equal to (38).

$$\begin{aligned}
&\left(\bar{x}^T(k)\otimes\bar{x}^T(k)\right)\mathrm{vec}(P_{i+1})\\
&-\beta\left(\bar{x}^T(k+1)\otimes\bar{x}^T(k+1)\right)\mathrm{vec}(P_{i+1})\\
&+\Big[2\bar{\alpha}\beta\left(\bar{x}^T(k)\otimes\left(\bar{K}_i\bar{x}(k)+u_a(k)\right)^T\right)\\
&+2\beta\left(1-\bar{\alpha}\right)\left(\bar{x}^T(k)\otimes u_a^T(k)\right)\Big]\mathrm{vec}(\bar{B}^TP_{i+1}\bar{A})\\
&-\Big[\bar{\alpha}\beta\left(\left(\bar{K}_i\bar{x}(k)-u_a(k)\right)^T\otimes\left(\bar{K}_i\bar{x}(k)+u_a(k)\right)^T\right)\\
&-\beta\left(1-\bar{\alpha}\right)\left(u_a^T(k)\otimes u_a^T(k)\right)\Big]\mathrm{vec}(\bar{B}^TP_{i+1}\bar{B})\\
=\;&\bar{x}^T(k)\bar{Q}\bar{x}(k)+\bar{\alpha}\bar{x}^T(k)\bar{K}_i^TR\bar{K}_i^T\bar{x}(k). \qquad (39)
\end{aligned}$$

To facilitate showing that (39) can be solved by using the least-square approach, define the following variables

$$\begin{aligned}
\mathcal{W}_l &= \left[\begin{array}{ccc}\mathcal{W}_{1,l} & \mathcal{W}_{2,l} & \mathcal{W}_{3,l}\end{array}\right],\\
\mathcal{W}_{1,l} &= \left(\bar{x}^T(k+l)\otimes\bar{x}^T(k+l)\right)\\
&\quad-\beta\left(\bar{x}^T(k+l+1)\otimes\bar{x}^T(k+l+1)\right),\\
\mathcal{W}_{2,l} &= 2\bar{\alpha}\beta\left(\bar{x}^T(k+l)\otimes\left(\bar{K}_i\bar{x}(k+l)+u_a(k+l)\right)^T\right)\\
&\quad+2\beta\left(1-\bar{\alpha}\right)\left(\bar{x}^T(k+l)\otimes u_a^T(k+l)\right),\\
\mathcal{W}_{3,l} &= -\bar{\alpha}\beta\left(\left(\bar{K}_i\bar{x}(k+l)-u_a(k+l)\right)^T\right.\\
&\quad\left.\otimes\left(\bar{K}_i\bar{x}(k+l)+u_a(k+l)\right)^T\right)\\
&\quad+\beta\left(1-\bar{\alpha}\right)\left(u_a^T(k+l)\otimes u_a^T(k+l)\right),\\
\Psi &= \left[\begin{array}{ccc}\mathrm{vec}(\Psi_1)^T & \mathrm{vec}(\Psi_2)^T & \mathrm{vec}(\Psi_3)^T\end{array}\right]^T,\\
\Psi_1 &= P_{i+1},\ \Psi_2=\bar{B}^TP_{i+1}\bar{A},\ \Psi_3=\bar{B}^TP_{i+1}\bar{B},\\
\Phi_l &= \bar{x}^T(k+l)\bar{Q}\bar{x}(k+l)\\
&\quad+\bar{\alpha}\bar{x}^T(k+l)\bar{K}_i^TR\bar{K}_i^T\bar{x}(k+l).
\end{aligned}$$

Then, (38) can be rewritten as

$$\mathcal{W}_l\Psi = \Phi_l,\ l=0.$$

Obviously, there exist $\varrho=(n_x+n_y)^2+n_u(n_x+n_y)+n_u^2$ unknown elements in Ψ. To use the least-square approach to solve those unknown elements, at least ϱ data should be collected. Define $\bar{\mathcal{W}}$ and $\bar{\Phi}$ as the collected data with

$$\begin{aligned}
\bar{\mathcal{W}} &= \left[\begin{array}{cccc}\mathcal{W}_0^T & \mathcal{W}_1^T & \ldots & \mathcal{W}_{\varrho-1}^T\end{array}\right]^T,\\
\bar{\Phi} &= \left[\begin{array}{cccc}\bar{\Phi}_0^T & \bar{\Phi}_1^T & \ldots & \bar{\Phi}_{\varrho-1}^T\end{array}\right]^T.
\end{aligned}$$

Combining the least-square approach, Ψ can be obtained as

$$\Psi = \left(\bar{\mathcal{W}}^T\bar{\mathcal{W}}\right)^{-1}\bar{\mathcal{W}}^T\bar{\Phi}.$$

Moreover, the optimal control gain can be represented as

$$\bar{K}_{i+1}=\left(R+\beta\Psi_3\right)^{-1}\beta\Psi_2.$$

Based on the above discussion, the off-policy learning control scheme without using exact system dynamics is presented in Algorithm 3.

Algorithm 3

1: Set the initial learning step $i=0$ and the learning error ϵ
2: Give an admissible control gain and $\bar{K}$ $u_a(k)=\bar{K}\bar{x}(k)+e(k)$ with $e(k)\neq 0$
3: Obtain $\bar{K}_{i+1}$, through solving (39) using the least-square approach
4: **if** $\|\bar{K}_{i+1}-\bar{K}_i\|<\epsilon$ **then**
5: Output $\bar{K}_{i+1}$ as the optimal control gain
6: **else**
7: $i=i+1$
8: Return to Step 3
9: **end if**

The core of Algorithm 3 is to iteratively solve (39). (39) is equivalent to solve the equation (32). In Theorem 7, the convergence of iteratively solving (32) has been proved and thus the convergence of Algorithm 3 can be also ensured.

To show the effectiveness of the proposed Algorithm 3, it is applied to an F16 aircraft system . The details are described in the following example.

Example 2: It is assumed that the physical system in Fig. 1 is a F16 aircraft system. The control signal is sent to the actuator through the cyber layer. The adversary can invade the cyber layer and prevent the control signal from successfully transmitting. The linear model of the F16 aircraft is described as [41], [42]

$$\dot{x}(t) = Ax(t)+Bu(t),$$

where $x=[\beta\ \omega\ \varphi]^T$ is the state with β being the angle of attack, ω being the pitch rate, and φ being the elevator deflection angle. $u(t)$ means the elevator actuator voltage, and

$$A=\left[\begin{array}{ccc}-1.01887 & 0.90506 & -0.00215\\ 0.82225 & -1.07741 & -0.17555\\ 0 & 0 & -1\end{array}\right],\ B=\left[\begin{array}{c}0\\0\\5\end{array}\right].$$

Under the sampling period 0.1, the parameters of the discrete-time model are as follows

$$A=\left[\begin{array}{ccc}0.9065 & 0.0816 & -0.0009\\ 0.0741 & 0.9012 & -0.0159\\ 0 & 0 & 0.9048\end{array}\right],\ B=\left[\begin{array}{c}-0.0002\\-0.0041\\0.4758\end{array}\right].$$

In this example, we assume the output is $y(k)=[1\ 0\ 0]x(k)$. The reference signal is a constant, that is $r(k+1)=r(k)$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3056633, IEEE Internet of Things Journal

To validate the proposed scheme, we set $Q = 1100$, $R = 1$, $\gamma = 0.8$, and $\bar{\alpha} = 0.6$. By directly solving the Riccati equation in (12), the optimal control gain can be obtained as

$$\bar{K} = \begin{bmatrix} -8.1286 & -4.2321 & 0.3088 & 13.8525 \end{bmatrix}.$$

The control objective is to make the state variable β to track the desired constant trajectory with the minimal cost. The simulation window is set as $[0,\ 300]$, and the initial state is given as $x(0) = [0.2\ 0.1\ 0.3]^T$. For the constant reference signal, it is defined as

$$r(k) = \begin{cases} 1, & 1 \le k \le 130, \\ 5, & 131 \le k \le 240, \\ 10, & 241 \le k \le 300. \end{cases}$$

Using the above optimal control gain, Fig. 9 depicts the responses of the system output and the reference signal under attacks. From the results, we can conclude that although attacks lead to some negative effects on the tracking performance, the designed control gain can be still effective.

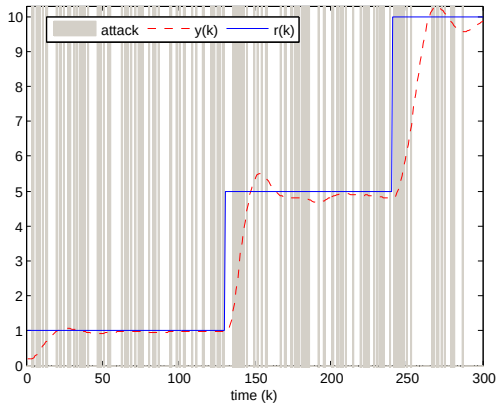

Fig. 9. Responses of the output signal $y(k)$ and reference signal $r(k)$ under DoS attacks.

In the following, Algorithm 3 is used to compute the optimal control gain. To this end, set $e(k) = randn\sin(9.8k) + \cos(10.2k)^2 + \sin(10k) + \cos(10k)$ as the probing noise with $randn$ being a function to generate a number from the standard normal distribution. Collect 35 data for each iteration. Fig. 10 shows the learning process of the gain $\bar{K}_i$. It is apparent that the optimal gain can be learned gradually. By recording each learning result, we can know that iterating 4 times yields the following optimal control gain with a satisfying precision

$$\bar{K}_4 = \begin{bmatrix} -8.1286 & -4.2321 & 0.3088 & 13.8525 \end{bmatrix}.$$

6.. Conclusion

The model-free optimal tracking control problem for CPS under DoS attacks has been solved in this paper. The system performance under attacks has been maintained by the proposed two learning based tracking control algorithms, the difference of which have been pointed out. Also, we have revealed the relation between the critical value of the successful attack probability and the existence and uniqueness

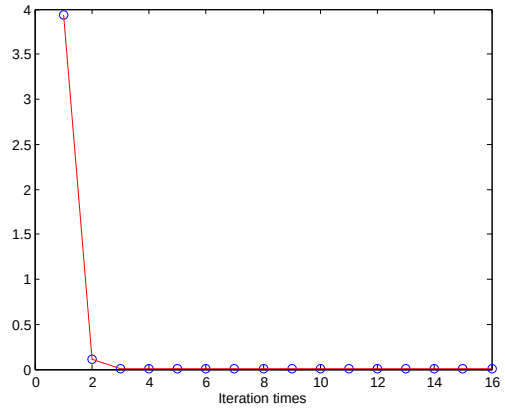

Fig. 10. Response of $\|\bar{K}_i - \bar{K}\|$.

of the solution to the Riccati equation. According to the learning approach used in this paper, the critical value can be obtained without using the exact system dynamics, which is different from those depending on the exact system dynamics, for example, [37], [38] and the references therein. Finally, the dc motor and the F16 aircraft systems have been used to evaluate the effectiveness of the proposed control schemes.

As we have discussed in the previous section, the learning technique used in this paper limits to the system structure. In the future, we will investigate how to develop learning based control scheme with stability guarantee for CPS, which is described by a Markov decision process.

References

[1] P. Antsaklis, "Goals and challenges in cyber-physical systems research editorial of the editor in chief," *IEEE Trans. Autom. Control*, vol. 59, no. 12, pp. 3117–3119, 2014.

[2] C. Wu, Z. Hu, J. Liu, and L. Wu, "Secure estimation for cyber-physical systems via sliding mode," *IEEE Trans. Cybern.*, vol. 48, no. 12, pp. 3420–3431, 2018.

[3] M. Pajic, I. Lee, and G. J. Pappas, "Attack-resilient state estimation for noisy dynamical systems," *IEEE Trans. Control Netw. Syst.*, vol. 4, no. 1, pp. 82–92, 2016.

[4] Z. Ju, H. Zhang, and Y. Tan, "Deception attack detection and estimation for a local vehicle in vehicle platooning based on a modified ufir estimator," *IEEE Internet Things J.*, vol. 7, no. 5, pp. 3693–3705, 2020.

[5] X. Wang, X. Luo, M. Zhang, Z. Jiang, and X. Guan, "Detection and isolation of false data injection attacks in smart grid via unknown input interval observer," *IEEE Internet Things J.*, vol. 7, no. 4, pp. 3214–3229, 2020.

[6] G. S. Paschos and L. Tassiulas, "Sustainability of service provisioning systems under stealth dos attacks," *IEEE Trans. Control Netw. Syst.*, vol. 4, no. 4, pp. 749–760, 2016.

[7] X. Fu and E. Modiano, "Fundamental limits of volume-based network dos attacks," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 3, no. 3, pp. 1–36, 2019.

[8] C. Wu, X. Li, W. Pan, J. Liu, and L. Wu, "Zero-sum game based optimal secure control under actuator attacks," *IEEE Trans. Autom. Control, DOI: 10.1109/TAC.2020.3029342*, 2020.

[9] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Trans. Autom. Control*, vol. 59, no. 6, pp. 1454–1467, 2014.

[10] Q. Hu, D. Fooladivanda, Y. H. Chang, and C. J. Tomlin, "Secure state estimation and control for cyber security of the nonlinear power systems," *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 3, pp. 1310–1321, 2017.

[11] X. Jin, W. M. Haddad, and T. Yucelen, "An adaptive control architecture for mitigating sensor and actuator attacks in cyber-physical systems," *IEEE Trans. Autom. Control*, vol. 62, no. 11, pp. 6058–6064, 2017.

[12] C. De Persis and P. Tesi, "Input-to-state stabilizing control under denial-of-service," *IEEE Trans. Autom. Control*, vol. 60, no. 11, pp. 2930–2944, 2015.

[13] D. Ding, Z. Wang, D. W. Ho, and G. Wei, "Observer-based event-triggering consensus control for multiagent systems with lossy sensors and cyber-attacks," *IEEE Trans. Cybern.*, vol. 47, no. 8, pp. 1936–1947, 2016.

[14] X.-M. Li, Q. Zhou, P. Li, H. Li, and R. Lu, "Event-triggered consensus control for multi-agent systems against false data-injection attacks," *IEEE Trans. Cybern.*, vol. 50, no. 5, pp. 1856–1866, 2019.

[15] Y. Li, D. Shi, and T. Chen, "False data injection attacks on networked control systems: A stackelberg game analysis," *IEEE Trans. Autom. Control*, vol. 63, no. 10, pp. 3503–3509, 2018.

[16] C. Wu, L. Wu, J. Liu, and Z.-P. Jiang, "Active defense based resilient sliding mode control under denial-of-service attacks," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 237–249, 2020.

[17] S. R. Etesami and T. Başar, "Dynamic games in cyber-physical security: An overview," *Dynamic Games and Applications*, pp. 1–30, 2019.

[18] E. Mousavinejad, X. Ge, Q.-L. Han, F. Yang, and L. Vlacic, "Resilient tracking control of networked control systems under cyber attacks," *IEEE Trans. Cybern., DOI: 10.1109/TCYB.2019.2948427*, 2019.

[19] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," *arXiv preprint arXiv:1801.01290*, 2018.

[20] Y. Jiang and Z.-P. Jiang, "Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics," *Automatica*, vol. 48, no. 10, pp. 2699–2704, 2012.

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

[22] W. Gao, J. Gao, K. Ozbay, and Z.-P. Jiang, "Reinforcement-learning-based cooperative adaptive cruise control of buses in the lincoln tunnel corridor with time-varying topology," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3796–3805, 2019.

[23] P. J. Werbos, "Neural networks for control and system identification," in *Proceedings of the 28th IEEE Conference on Decision and Control,*, pp. 260–265, IEEE, 1989.

[24] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT press, 2018.

[25] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Model-free Q-learning designs for linear discrete-time zero-sum games with application to H-infinity control," *Automatica*, vol. 43, no. 3, pp. 473–481, 2007.

[26] S. A. A. Rizvi and Z. Lin, "Output feedback Q-learning for discrete-time linear zero-sum games with application to the H_∞ control," *Automatica*, vol. 95, pp. 213–221, 2018.

[27] B. Pang, Z.-P. Jiang, and I. Mareels, "Reinforcement learning for adaptive optimal control of continuous-time linear periodic systems," *Automatica*, vol. 118, p. 109035, 2020.

[28] H. Xu, S. Jagannathan, and F. L. Lewis, "Stochastic optimal control of unknown linear networked control system in the presence of random delays and packet losses," *Automatica*, vol. 48, no. 6, pp. 1017–1030, 2012.

[29] B. Kiumarsi, F. L. Lewis, H. Modares, 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, 2014.

[30] 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 Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4607–4620, 2017.

[31] H. Ma, H. Li, R. Lu, and T. Huang, "Adaptive event-triggered control for a class of nonlinear systems with periodic disturbances," *Science China Information Sciences*, vol. 63, pp. 1–15, 2020.

[32] Y. Ni, Z. Guo, Y. Mo, and L. Shi, "On the performance analysis of reset attack in cyber-physical systems," *IEEE Trans. Autom. Control, DOI: 10.1109/TAC.2019.2914655*, 2019.

[33] P. Griffioen, S. Weerakkody, and B. Sinopoli, "An optimal design of a moving target defense for attack detection in control systems," in *American Control Conference*, pp. 4527–4534, IEEE, 2019.

[34] Y. Li, A. S. Mehr, and T. Chen, "Multi-sensor transmission power control for remote estimation through a SINR-based communication channel," *Automatica*, vol. 101, pp. 78–86, 2019.

[35] Y. Li, D. E. Quevedo, S. Dey, and L. Shi, "SINR-based DoS attack on remote state estimation: A game-theoretic approach," *IEEE Trans. Control Netw. Syst.*, vol. 4, no. 3, pp. 632–642, 2016.

[36] M. Salehi and J. Proakis, "Digital communications," *McGraw-Hill Education*, 2007.

[37] Y. Mo, E. Garone, and B. Sinopoli, "LQG control with Markovian packet loss," in *European Control Conference*, pp. 2380–2385, IEEE, 2013.

[38] B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M. I. Jordan, and S. S. Sastry, "Kalman filtering with intermittent observations," *IEEE Trans. Autom. Control*, vol. 49, no. 9, pp. 1453–1464, 2004.

[39] D. J. Tylavsky and G. R. Sohie, "Generalization of the matrix inversion lemma," *Proc. IEEE*, vol. 74, no. 7, pp. 1050–1052, 1986.

[40] Y. Shi, J. Huang, and B. Yu, "Robust tracking control of networked control systems: application to a networked DC motor," *IEEE Trans. Ind. Electron.*, vol. 60, no. 12, pp. 5864–5874, 2012.

[41] H. Modares, F. L. Lewis, and Z.-P. Jiang, "H_∞ tracking control of completely unknown continuous-time systems via off-policy reinforcement learning," *IEEE Trans. Control Netw. Syst.*, vol. 26, no. 10, pp. 2550–2562, 2015.

[42] L. T. Nguyen, *Simulator study of stall/post-stall characteristics of a fighter airplane with relaxed longitudinal static stability*. National Aeronautics and Space Administration, 1979.

Chengwei Wu received the B.S. degree in management from the Arts and Science College, Bohai University, Jinzhou, China, in 2013, and the M.S. degree from Bohai University, in 2016. From July 2015 to December 2015, he was a Research Assistant in the Department of Mechanical Engineering, The Hong Kong Polytechnic University. He is currently pursuing the Ph.D. degree with the Harbin Institute of Technology, Harbin, China. His research interests include sliding mode control, reinforcement learning and networked control systems.

Wei Pan received the Ph.D. degree in Bioengineering from Imperial College London in 2016. He is currently an Assistant Professor at Department of Cognitive Robotics, Delft University of Technology. Until May 2018, he was a Project Leader at DJI, Shenzhen, China, responsible for machine learning research for DJI drones and AI accelerator. He is the recipient of Dorothy Hodgkins Postgraduate Awards, Microsoft Research Ph.D. Scholarship and Chinese Government Award for Outstanding Students Abroad, Shenzhen Peacock Plan Award. He is an active reviewer and committee member for many international journals and conferences. His research interests include machine learning and control theory with applications in robotics.

Guanghui Sun received the B.S. degree in automation and the M.S. and Ph.D. degrees in control science and engineering from Harbin Institute of Technology, Harbin, China, in 2005, 2007, and 2010, respectively. He is currently a Professor in the Department of Control Science and Engineering, Harbin Institute of Technology. His research interests include fractional-order systems, networked control systems, and sliding mode control.

Jianxing Liu received the B.S. degree in mechanical engineering in 2008, the M.E. degree in control science and engineering in 2010, both from Harbin Institute of Technology, Harbin, China and the Ph.D. degree in Automation from the Technical University of Belfort-Montbeliard (UTBM), France, in 2014. Since 2014, he joined Harbin Institute of Technology, Harbin, China. His current research interests include nonlinear control algorithms, sliding mode control, and their applications in industrial electronics systems and renewable energy systems.

Ligang Wu (M'10-SM'12-F'19) received the B.S. degree in Automation from Harbin University of Science and Technology, China in 2001; the M.E. degree in Navigation Guidance and Control from Harbin Institute of Technology, China in 2003; the Ph.D. degree in Control Theory and Control Engineering from Harbin Institute of Technology, China in 2006. From January 2006 to April 2007, he was a Research Associate in the Department of Mechanical Engineering, The University of Hong Kong, Hong Kong. From September 2007 to June 2008, he was a Senior Research Associate in the Department of Mathematics, City University of Hong Kong, Hong Kong. From December 2012 to December 2013, he was a Research Associate in the Department of Electrical and Electronic Engineering, Imperial College London, London, UK. In 2008, he joined the Harbin Institute of Technology, China, as an Associate Professor, and was then promoted to a Full Professor in 2012. Prof. Wu was the winner of the National Science Fund for Distinguished Young Scholars in 2015, and received China Young Five Four Medal in 2016. He was named as the Distinguished Professor of Chang Jiang Scholar in 2017, and was named as the Highly Cited Researcher in 2015-2019.

Prof. Wu currently serves as an Associate Editor for a number of journals, including IEEE TRANSACTIONS ON AUTOMATIC CONTROL, IEEE/ASME TRANSACTIONS ON MECHATRONICS, IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, *Information Sciences, Signal Processing,* and *IET Control Theory and Applications.* He is an Associate Editor for the Conference Editorial Board, IEEE Control Systems Society. He is also a Fellow of IEEE. Prof. Wu has published 7 research monographs and more than 170 research papers in international referred journals. His current research interests include switched systems, stochastic systems, computational and intelligent systems, sliding mode control, and advanced control techniques for power electronic systems.