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

Optimal Tracking Control of Linear Discrete-Time Systems Under Cyber Attacks ⋆

Hao Liu, * **Hui Qiu**

* *School of Automation, Shenyang Aerospace University, Shenyang, 110136, China (e-mail: lh_hit_1985@163.com).*

Abstract: In this paper, the optimal tracking control problem is solved based on the reinforcement learning for linear systems subject to multiple false-data-injection (FDI) attacks. An augmented system is established, which includes the original system and reference-trajectory generator system. The corresponding optimal control issue is formulated as a game problem between the system and malicious adversaries. A Q-learning algorithm is proposed to solve the game algebraic Riccati equation without requiring any knowledge about the dynamics of the augmented system. Finally, an example is provided to show that the system output can track the reference trajectory under cyber attacks.

Keywords: Linear systems, optimal tracking control; false-data-injection attacks; game theory; Q-learning.

1. INTRODUCTION

Optimal tracking control is a main research field in control theory, and the objective is to design an optimal controller such that the system output can track a reference trajectory in an optimal sense. This can be achieved by minimizing a predefined quadratic performance index. It should be noted that the communication is vulnerable to attacks launched by malicious adversaries when the controller and the plant are linked through wireless network. Thus, it is meaningful to consider the effect of cyber-attacks.

In Teixeira et al. (2015), cyber-secure networked control is modeled and analyzed, and the attack space is divided into three dimensions: the adversary's model knowledge, disclosure and disruption resources. Consequently, the corresponding attacks are mainly classified into three types, i.e., denial-of-service attacks Persis et al. (2016); Qin et al. (2018); Ding et al. (2017), false-data-injection (FDI) attacks Kung et al. (2017); Bai et al. (2017); Hu et al. (2018) and replay attacks Mo et al. (2009); Zhu et al. (2014); Chen et al. (2018). FDI attacks aim to replace the original data with false data injected by adversaries. Note that these injected data may deteriorate the system performance. Moreover, FDI attacks require to know the model knowledge and disruption resources.

The security problem for a networked control system was investigated in Hu et al. (2018), and a new necessary and sufficient condition for the insecurity was derived. A notion of ϵ-stealthiness was given in Bai et al. (2017) to quantify the detectability of attacks in stochastic cyber-physical systems. Based on the entropy theory, the performance degradation in the presence of attacks was characterized. It is noted that a Stackelberg game theory can be used to deal with FDI attacks if the two players make their decisions sequentially. In Li et al. (2018), the interactive decision-making process between the defender and the attacker were studied in a Stackelberg game framework, and the optimal strategies for both sides were solved by using linear programming approach.

It should be emphasized that, however, the design of controller is rarely considered to defend FDI attacks. For instance, the controller design problem of CPS was studied in Ye et al. (2019) to ensure the reliability and security when actuator faults in physical layers and attacks in cyber layers occur simultaneously. In Ding et al. (2017), the consensus control problem was investigated for a class of multiagent systems with lossy sensors and cyber-attacks, in which a dynamic output feedback controller was designed.

As a result, optimal tracking control for discrete-time systems under multiple FDI attacks are investigated in this work. It is assumed that the controller can acquire the information of the plant state and the reference signal. Moreover, the corresponding issue is formulated into a game problem, which can be solved by using reinforcement learning method.

Notations: The superscript T stands for matrix transposition; $\mathbb{R}^n$ denotes the n-dimensional Euclidean space; I and 0 represent the identity matrix and the zero matrix, respectively; the notation $P > 0$ means that P is real symmetric and positive definite; $A \otimes B$ is the Kronecker product of A and B.

2. PROBLEM FORMULATION

Consider the following discrete-time linear systems

⋆ This work is supported by the National Natural Science Foundation of China (61703286) and the Project from the Education Department of Liaoning Province (JYT19039).

$$x_{k+1} = Ax_k + Bu_k \tag{1}$$
$$y_k = Cx_k \tag{2}$$

where $x_k \in \mathbb{R}^n$ is the system state, $u_k \in \mathbb{R}^m$ is the control input, and $y_k \in \mathbb{R}^p$ is the system output. Assume that (A, B) and (A, C) are controllable and observable pairs, respectively.

In this paper, it is assumed that the reference input satisfies the following model

$$y^r_{k+1} = Ty^r_k \tag{3}$$

where $y^r_k \in \mathbb{R}^p$ and T needs not to be Hurwitz. The aim of controller is to make the system output y_k track the reference input y^r_k. Moreover, we assume that the plant and the controller are linked through wireless network. Assume that the wireless network communication can be attacked by adversaries and all attackers have full knowledge of the system. The communication between the remote controller and the system can be eavesdropped and attacked, and the modified data u^a_k can be represented by

$$u^a_k = u_k + \sum_{j=1}^{q} \Gamma^j a^j_k \tag{4}$$

where q is the number of adversaries, $a^j_k \in \mathbb{R}^m$, $j \in \mathcal{Q} = \{1, 2, \ldots, q\}$, is the false-data injected by attacker j at time step k, and matrix Γ^j satisfies

$$\Gamma^j = diag\{\beta^j_1, \beta^j_2, \cdots, \beta^j_m\} \tag{5}$$

where $\beta^j_i \in \{0, 1\}$, $i = 1, 2, \ldots, m$, $j \in \mathcal{Q}$. $\beta^j_i = 1$ means that i^{th} communication channel is attacked by j^{th} attacker; otherwise, the corresponding channel is not injected false data. It should be pointed out that Γ^j can be determined by using different method, such as game theory approach, assuming β^j_i satisfies bernoulli distribution, etc. In this paper, it is assumed that the matrix Γ^j is known in advance.

Considering modified control input (4), the equation (1) becomes

$$x_{k+1} = Ax_k + Bu^a_k = Ax_k + Bu_k + \sum_{j=1}^{q} B\Gamma^j a^j_k \tag{6}$$

Now, define the following tracking error

$$e_k = y_k - y^r_k = Cx_k - y^r_k \tag{7}$$

Based on (6) and the reference input dynamics (3), the following augmented system can be constructed

$$\bar{x}_{k+1} = \begin{bmatrix} A & 0 \\ 0 & T \end{bmatrix} \bar{x}_k + \begin{bmatrix} B \\ 0 \end{bmatrix} u_k + \sum_{j=1}^{q} \begin{bmatrix} B\Gamma^j \\ 0 \end{bmatrix} a^j_k$$
$$= \bar{A}\bar{x}_k + \bar{B}u_k + \sum_{j=1}^{q} \bar{\Gamma}^j a^j_k \tag{8}$$

where $\bar{x}_k = \left[x^T_k,\ (y^r_k)^T\right]^T$. Then, the tracking error e_k can be given by

$$e_k = [C\ \ -I]\bar{x}_k = \bar{C}\bar{x}_k \tag{9}$$

Note that the controller can be designed in the form of different types, such as state feedback, dynamic output feedback, etc. On the other hand, the false-data injected by adversaries can also be given in many different types, which means that an attacker can utilizes some different types of signals, such as the system states, the reference input, and so on. In this paper, however, the control input u_k and the injected false data a^j_k are assumed to be linear functions of x_k and y^r_k, which can be represented as follows

$$u_k = \mu(x_k, y^r_k) = K_1 x_k + K_2 y^r_k = K\bar{x}_k \tag{10}$$

and

$$a^j_k = h^j(x_k, y^r_k) = L^j_1 x_k + L^j_2 y^r_k = L^j \bar{x}_k \tag{11}$$

where $j \in \mathcal{Q}$, K and L^j are constant feedback gain matrices to be determined. The advantage of such design can be seen in the following analysis.

The objective of the defender, i.e., the controller, is to minimize the following reward function at time step k

$$\mathcal{R}_d(x_k, y^r_k, u_k) = \sum_{i=k}^{\infty} \gamma^{i-k} \left(e^T_i Q_e e_i + u^T_i R u_i\right) \tag{12}$$

where $Q_e > 0$, $R > 0$ and $\gamma \in (0, 1)$ is the discount factor. It is noted that $\gamma = 1$ can be chosen if matrix T is Hurwitz. Therefore, the optimal control policy can be obtained by solving

$$u^*_k = \arg\min_{u_k} \mathcal{R}_d(x_k, y^r_k, u_k) \tag{13}$$

Similarly, the reward for the attacker j at time step k is

$$\mathcal{R}_a(x_k, y^r_k, a^j_k) = \sum_{i=k}^{\infty} \gamma^{i-k} \left(e^T_i Q_e e_i - \vartheta^j (a^j_i)^T (a^j_i)\right) \tag{14}$$

where $\vartheta^j > 0$ is a pre-defined weighting parameter, and it is assumed that both players know the reward functions of their opponents. The optimal attack policy for j^{th} attacker can be given by

$$a^{j,*}_k = \arg\max_{a^j_k} \mathcal{R}_a(x_k, y^r_k, a^j_k), \quad j \in \mathcal{Q} \tag{15}$$

Remark 1. Equations (13) and (15) show that the defender aims to reduce the tracking error and consume less energy; while the purpose of j^{th} attacker is to increase the tracking error. Therefore, the objectives of both players are opposite.

3. MAIN RESULTS

Define the following functions

$$\begin{aligned} &J(\bar{x}_k, u_k, a^1_k, \cdots, a^q_k) \\ =&J(x_k, y^r_k, u_k, a^1_k, \cdots, a^q_k) \\ =&\sum_{i=k}^{\infty} \gamma^{i-k} (e^T_i Q_e e_i + u^T_i R u_i - \sum_{j=1}^{q} \vartheta^j (a^j_i)^T (a^j_i)) \end{aligned} \tag{16}$$

$$\bar{J} = \inf_{u_k} \sup_{a^1_k, \cdots, a^q_k} J(\bar{x}_k, u_k, a^1_k, \cdots, a^q_k) \tag{17}$$

$$\underline{J} = \sup_{a^1_k, \cdots, a^q_k} \inf_{u_k} J(\bar{x}_k, u_k, a^1_k, \cdots, a^q_k) \tag{18}$$

If $J^* = \bar{J} = \underline{J}$, then J^* is called the value of the corresponding game and $(u^*_k, a^{1,*}_k, \cdots, a^{q,*}_k)$ is the Nash equilibrium at time step k, namely, the saddle-point solution.

Then, calculating (13) and (15) is equivalent to find the saddle point $(u^*_k, a^{1,*}_k, \cdots, a^{q,*}_k)$ such that

$$\begin{aligned} J^*(x_k, y^r_k) &= \min_{u_k} \max_{a^1_k, \cdots, a^q_k} J(x_k, y^r_k, u_k, a^1_k, \cdots, a^q_k) \\ &= \max_{a^1_k, \cdots, a^q_k} \min_{u_k} J(x_k, y^r_k, u_k, a^1_k, \cdots, a^q_k) \end{aligned} \tag{19}$$

Define the following utility function at time step k

$$\begin{aligned} r_k &= r_k(x_k, y_k^r, u_k, a_k^1, \cdots, a_k^q) \\ &= \bar{x}_k^T Q_x \bar{x}_k + u_k^T R u_k - \sum_{j=1}^{q} \vartheta^j (a_k^j)^T (a_k^j) \end{aligned} \tag{20}$$

where $Q_x = \bar{C}^T Q_e \bar{C}$. Considering the state feedback control, then one can obtain from (16) that

$$\begin{aligned} J(\bar{x}_k) &= J(\bar{x}_k, \mu(\bar{x}_k), h^1(\bar{x}_k), \cdots, h^q(\bar{x}_k)) \\ &= \sum_{i=k}^{\infty} \gamma^{i-k} r_k(\bar{x}_k, \mu(\bar{x}_k), h^1(\bar{x}_k), \cdots, h^q(\bar{x}_k)) \end{aligned} \tag{21}$$

Through a simple calculation, (21) can be re-written as

$$\begin{aligned} J(\bar{x}_k) =& r_k(\bar{x}_k, \mu(\bar{x}_k), h^1(\bar{x}_k), \cdots, h^q(\bar{x}_k)) + \gamma J(\bar{x}_{k+1}) \\ =& \bar{x}_k^T Q_x \bar{x}_k + u_k^T R u_k - \sum_{j=1}^{q} \vartheta^j (a_k^j)^T (a_k^j) + \gamma J(\bar{x}_{k+1}) \end{aligned} \tag{22}$$

which is the Bellman equation of optimal control. According to the Bellman optimality principle, the value function $J(\bar{x}_k)$ is quadratic in the state $\bar{x}_k$ at time step k, and can be represented as

$$J(\bar{x}_k) = \bar{x}_k^T P \bar{x}_k \tag{23}$$

where $P \in \mathbb{R}^{(n+p)\times(n+p)}$ is positive definite and to be determined. Then, the Bellman equation (22) becomes

$$\begin{aligned} \bar{x}_k^T P \bar{x}_k =& \bar{x}_k^T Q_x \bar{x}_k + u_k^T R u_k - \sum_{j=1}^{q} \vartheta^j (a_k^j)^T (a_k^j) \\ &+ \gamma \bar{x}_{k+1}^T P \bar{x}_{k+1} \end{aligned} \tag{24}$$

For simplicity, we can define $\vartheta = diag\{\vartheta^1, \vartheta^2, \cdots, \vartheta^q\}$, $\bar{\Gamma} = [\bar{\Gamma}^1\ \bar{\Gamma}^2\ \cdots\ \bar{\Gamma}^q]$ and $a_k = [(a_k^1)^T\ (a_k^2)^T\ \cdots\ (a_k^q)^T]^T$. Therefore, the optimal strategies satisfy

$$\begin{aligned} &\begin{bmatrix} R + \gamma \bar{B}^T P \bar{B} & \gamma \bar{B}^T P \bar{\Gamma} \\ \gamma \bar{\Gamma}^T P \bar{B} & \gamma \bar{\Gamma}^T P \bar{\Gamma} - \vartheta \otimes I \end{bmatrix} \begin{bmatrix} u_k \\ a_k \end{bmatrix} \\ &= - \begin{bmatrix} \gamma \bar{B}^T P \bar{A} \\ \gamma \bar{\Gamma}^T P \bar{A} \end{bmatrix} \bar{x}_k \end{aligned} \tag{25}$$

The following optimal policies for both players can be obtained by solving (25)

$$u_k^* = K(P)\bar{x}_k = \left(R + \gamma \bar{B}^T P \bar{B} - \Omega \bar{B}\right)^{-1} \Omega \bar{A} \bar{x}_k \tag{26}$$

and

$$a_k^* = L(P)\bar{x}_k = -(\Theta \bar{\Gamma} - \vartheta \otimes I)^{-1} \Theta \bar{A} \bar{x}_k \tag{27}$$

where

$$\begin{aligned} \Omega &= \gamma^2 \bar{B}^T P \bar{\Gamma} \left(\gamma \bar{\Gamma}^T P \bar{\Gamma} - \vartheta \otimes I\right)^{-1} \bar{\Gamma}^T P \\ \Theta &= \left[(\Theta^1)^T\ (\Theta^2)^T\ \cdots\ (\Theta^q)^T\right]^T \\ \Theta^j &= \gamma (\bar{\Gamma}^j)^T P \left[I - \gamma \bar{B}(R + \gamma \bar{B}^T P \bar{B})^{-1} \bar{B}^T P\right] \\ L(P) &= \left[(L^1(P))^T\ (L^2(P))^T\ \cdots (L^q(P))^T\right]^T \end{aligned}$$

and P satisfies the following game Riccati equation

$$\begin{aligned} P =& Q_x + \gamma \bar{A}^T P \bar{A} - \gamma^2 \begin{bmatrix} \bar{A}^T P \bar{B} & \bar{A}^T P \bar{\Gamma} \end{bmatrix} \\ &\times \begin{bmatrix} R + \gamma \bar{B}^T P \bar{B} & \gamma \bar{B}^T P \bar{\Gamma} \\ \gamma \bar{\Gamma}^T P \bar{B} & \gamma \bar{\Gamma}^T P \bar{\Gamma} - \vartheta \otimes I \end{bmatrix}^{-1} \begin{bmatrix} \bar{B}^T P \bar{A} \\ \bar{\Gamma}^T P \bar{A} \end{bmatrix} \end{aligned} \tag{28}$$

In order to obtain the optimal strategies for both sides, the knowledge of system dynamics (A, B, Γ, T) is needed. In this paper, a Q-function is given for the game between the defender and attackers, which does not need to know (A, B, Γ, T). Define the following Q-function Bellman equation

$$\begin{aligned} &Q(\bar{x}_k, u_k, a_k^1, \cdots, a_k^q) \\ &= r_k(\bar{x}_k, u_k, a_k^1, \cdots, a_k^q) + \gamma J(\bar{x}_{k+1}) \\ &= \bar{x}_k^T Q_x \bar{x}_k + u_k^T R u_k - \sum_{j=1}^{q} \vartheta^j (a_k^j)^T (a_k^j) + \gamma J(\bar{x}_{k+1}) \\ &= \bar{x}_k^T Q_x \bar{x}_k + u_k^T R u_k - a_k^T (\vartheta \otimes I) a_k \\ &\quad + \gamma Q(\bar{x}_{k+1}, u_{k+1}, a_{k+1}^1, \cdots, a_{k+1}^q) \end{aligned} \tag{29}$$

which can be re-written as the following compact form

$$\begin{aligned} &Q(\bar{x}_k, u_k, a_k^1, \cdots, a_k^q) = Q(\bar{x}_k, u_k, a_k) \\ &= \eta_k^T \begin{bmatrix} H_{\bar{x}\bar{x}} & H_{\bar{x}u} & H_{\bar{x}a} \\ H_{u\bar{x}} & H_{uu} & H_{ua} \\ H_{a\bar{x}} & H_{au} & H_{aa} \end{bmatrix} \eta_k = \eta_k^T H \eta_k \end{aligned} \tag{30}$$

where

$$\begin{aligned} &H = H^T,\ \eta_k = [\bar{x}_k\ u_k\ a_k]^T = \left[\bar{x}_k\ u_k\ a_k^1\ \cdots\ a_k^q\right]^T, \\ &H_{\bar{x}a} = \left[H_{\bar{x}a^1}\ H_{\bar{x}a^2}\ \cdots\ H_{\bar{x}a^q}\right],\ H_{\bar{x}a^j} = \gamma \bar{A}^T P \bar{\Gamma}^j, \\ &H_{ua} = \left[H_{ua^1}\ H_{ua^2}\ \cdots\ H_{ua^q}\right],\ H_{ua^j} = \gamma \bar{B}^T P \bar{\Gamma}^j, \\ &H_{aa} = \begin{bmatrix} H_{a^1a^1} & \cdots & H_{a^1a^q} \\ \vdots & \ddots & \vdots \\ H_{a^qa^1} & \cdots & H_{a^qa^q} \end{bmatrix}, \\ &H_{a^ia^j} = \begin{cases} \gamma(\bar{\Gamma}^i)^T P (\bar{\Gamma}^j), & i \neq j \\ \gamma(\bar{\Gamma}^j)^T P (\bar{\Gamma}^j) - \vartheta^j \otimes I, & i = j \end{cases}. \end{aligned}$$

The optimal policies for both the defender and attackers can be obtained by solving the following equations for u_k and a_k^j.

$$\frac{\partial Q(\bar{x}_k, u_k, a_k^1, \cdots, a_k^q)}{\partial u_k} = 0 \tag{31}$$

$$\frac{\partial Q(\bar{x}_k, u_k, a_k^1, \cdots, a_k^q)}{\partial a_k^j} = 0, \quad j \in \mathcal{Q} \tag{32}$$

which yields

$$\begin{aligned} u_k^* =& \left(H_{uu} - H_{ua} H_{aa}^{-1} H_{au}\right)^{-1} \\ &\times \left(H_{ua} H_{aa}^{-1} H_{a\bar{x}} - H_{u\bar{x}}\right) \bar{x}_k \end{aligned} \tag{33}$$

and

$$\begin{aligned} a_k^* =& \left(H_{aa} - H_{au} H_{uu}^{-1} H_{ua}\right)^{-1} \\ &\times \left(H_{au} H_{uu}^{-1} H_{u\bar{x}} - H_{a\bar{x}}\right) \bar{x}_k \end{aligned} \tag{34}$$

where $a_k^* = \left[(a_k^{1,*})^T\ (a_k^{2,*})^T\ \cdots\ (a_k^{q,*})^T\right]^T$.

Next, the aim is to express the Q-function given in (30) in terms of input-output data instead of the system state $\bar{x}_k$.

Lemma 2. Assume the system (2) and (8) is observable. Then, the system state can be re-written in terms of measured input/output sequences as

$$\begin{aligned} \bar{x}_k =& M_e \bar{e}_{k-1,k-N} + M_u \bar{u}_{k-1,k-N} \\ &+ \sum_{j=1}^{q} M_a^j \bar{a}_{k-1,k-N}^j \end{aligned} \tag{35}$$

where

$$\bar{e}_{k-1,k-N} = \left[e_{k-1}^T\ e_{k-2}^T\ \cdots\ e_{k-N}^T\right]^T,$$
$$\bar{u}_{k-1,k-N} = \left[u_{k-1}^T\ u_{k-2}^T\ \cdots\ u_{k-N}^T\right]^T,$$
$$\bar{a}_{k-1,k-N}^j = \left[(a_{k-1}^j)^T\ (a_{k-2}^j)^T\ \cdots\ (a_{k-N}^j)^T\right]^T,$$
$$M_e = \bar{A}^N \left(V_N^T V_N\right)^{-1} V_N^T,$$
$$M_u = U_N^1 - M_e W_N^1, \quad M_a^{j-1} = U_N^j - M_e W_N^j,$$
$$V_N = \left[(\bar{C}\bar{A}^{N-1})^T\ (\bar{C}\bar{A}^{N-2})^T\ \cdots\ (\bar{C}\bar{A})^T\ \bar{C}^T\right]^T,$$
$$U_N^1 = \left[\bar{B}\ \bar{A}\bar{B}\ \cdots\ \bar{A}^{N-1}\bar{B}\right],$$
$$U_N^j = \left[\bar{\Gamma}^{j-1}\ \bar{A}\bar{\Gamma}^{j-1}\ \cdots\ \bar{A}^{N-1}\bar{\Gamma}^{j-1}\right],$$
$$W_N^1 = \begin{bmatrix} 0 & \bar{C}\bar{B} & \bar{C}\bar{A}\bar{B} & \cdots & \bar{C}\bar{A}^{N-2}\bar{B} \\ 0 & 0 & \bar{C}\bar{B} & \cdots & \bar{C}\bar{A}^{N-3}\bar{B} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \bar{C}\bar{B} \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix},$$
$$W_N^j = \begin{bmatrix} 0 & \bar{C}\bar{\Gamma}^{j-1} & \bar{C}\bar{A}\bar{\Gamma}^{j-1} & \cdots & \bar{C}\bar{A}^{N-2}\bar{\Gamma}^{j-1} \\ 0 & 0 & \bar{C}\bar{\Gamma}^{j-1} & \cdots & \bar{C}\bar{A}^{N-3}\bar{\Gamma}^{j-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \bar{C}\bar{\Gamma}^{j-1} \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix},$$
$j = 2, 3, \ldots, q+1.$

Proof. The proof can be easily completed and omitted here due to the space limitation.

It should be pointed out that there exists a constant κ such that $rank(V_N) < n+p$ for $N < \kappa$ and that $rank(V_N) = n+p$ for $N \geq \kappa$, where κ is the observability index. Consequently, one can choose $N \geq \kappa$, and V_N has full column rank $n+p$. In addition, $\bar{e}_{k-1,k-N}$, $\bar{u}_{k-1,k-N}$ and $\bar{a}_{k-1,k-N}^j$ are the tracking error output, control input and j^{th} false-data input, respectively. V_N is the observability matrix, and U_N^i, $i = 1, 2, \ldots, q+1$ are the controllability matrices with respect to control and attacks. W_N^i, $i = 1, 2, \ldots, q+1$ are the Toeplitz matrices. It can be seen from Lemma 1 that the augmented system state $\bar{x}_k$ is represented as the measured input-output sequences. Next, the Q-function (30) will be described by the same input-output sequences.

Define the following vector

$$\xi_k = \begin{bmatrix} \bar{e}_{k-1,k-N} \\ \bar{u}_{k-1,k-N} \\ \bar{a}_{k-1,k-N} \\ u_k \\ a_k \end{bmatrix} \in \mathbb{R}^{[(m+1)p+m]N+(p+1)m} \tag{36}$$

Then, the Q-function can be re-written in the form of (37), where $M_a = \left[M_a^1\ M_a^2\ \cdots\ M_a^q\right]$. As a result, the optimal defender's policy u_k^* and attackers' policy a_k^* can be obtained by solving (31) and (32) simultaneously, where $Q(\bar{x}_k, u_k, a_k) = Q(\xi_k)$, which yields

$$u_k^* = \left(\bar{H}_{uu} - \bar{H}_{ua}(\bar{H}_{aa})^{-1}\bar{H}_{au}\right)^{-1} \times \left(\bar{H}_{ua}(\bar{H}_{aa})^{-1}\phi_{k-1,k-N} - \psi_{k-1,k-N}\right) \tag{38}$$

and

$$a_k^* = \left(\bar{H}_{aa} - \bar{H}_{au}(\bar{H}_{uu})^{-1}\bar{H}_{ua}\right)^{-1} \times \left(\bar{H}_{au}(\bar{H}_{uu})^{-1}\psi_{k-1,k-N} - \phi_{k-1,k-N}\right) \tag{39}$$

where

$$\begin{aligned}\phi_{k-1,k-N} =& \bar{H}_{a\bar{e}}\bar{e}_{k-1,k-N} + \bar{H}_{a\bar{u}}\bar{u}_{k-1,k-N} \\ &+ \bar{H}_{a\bar{a}}\bar{a}_{k-1,k-N}, \\ \psi_{k-1,k-N} =& \bar{H}_{u\bar{e}}\bar{e}_{k-1,k-N} + \bar{H}_{u\bar{u}}\bar{u}_{k-1,k-N} \\ &+ \bar{H}_{u\bar{a}}\bar{a}_{k-1,k-N}, \\ \bar{a}_{k-1,k-N} =& \left[(\bar{a}_{k-1,k-N}^1)^T\ \cdots\ (\bar{a}_{k-1,k-N}^q)^T\right]^T.\end{aligned}$$

The following theorem can be obtained immediately from the above analysis, and the proof is omitted here.

Theorem 3. Assume the augmented systems (8)-(9) are observable. The optimal policies for both defender and attackers can be calculated according to (38) and (39), respectively, which are functions of measured input/output sequences and independent of the system state.

By combining (29) and (37), the input/output form of Bellman equation for Q-function can be re-written as

$$\begin{aligned}\xi_k^T \bar{H} \xi_k =& \bar{x}_k^T Q_x \bar{x}_k + u_k^T R u_k - a_k^T(\vartheta \otimes I)a_k \\ &+ \gamma \xi_{k+1}^T \bar{H} \xi_{k+1}\end{aligned} \tag{40}$$

where u_{k+1} and a_{k+1} can be calculated by

$$u_{k+1} = \left(\bar{H}_{uu} - \bar{H}_{ua}(\bar{H}_{aa})^{-1}\bar{H}_{au}\right)^{-1} \times \left(\bar{H}_{ua}(\bar{H}_{aa})^{-1}\phi_{k,k-N+1} - \psi_{k,k-N+1}\right)$$

and

$$a_{k+1} = \left(\bar{H}_{aa} - \bar{H}_{au}(\bar{H}_{uu})^{-1}\bar{H}_{ua}\right)^{-1} \times \left(\bar{H}_{au}(\bar{H}_{uu})^{-1}\psi_{k,k-N+1} - \phi_{k,k-N+1}\right)$$

Now, we linearly parameterize the Q-function as follows

$$\begin{aligned}\xi_k^T \bar{H} \xi_k =& \bar{H}_{11}(\xi_k^1)^2 + 2\bar{H}_{12}\xi_k^1\xi_k^2 + \cdots + 2\bar{H}_{1\ell}\xi_k^1\xi_k^\ell \\ &+ \bar{H}_{22}(\xi_k^2)^2 + 2\bar{H}_{23}\xi_k^2\xi_k^3 + \cdots 2\bar{H}_{2\ell}\xi_k^2\xi_k^\ell \\ &+ \cdots + \bar{H}_{\ell\ell}(\xi_k^\ell)^2 \\ =& h^T \bar{\xi}_k\end{aligned} \tag{41}$$

where

$$\bar{H} = \bar{H}^T,\ \xi_k = \left[\xi_k^1\ \xi_k^2 \cdots \xi_k^\ell\right]^T,$$
$$\ell = [(m+1)p+m]N + (p+1)m,$$
$$h = \left[\bar{H}_{11}\ 2\bar{H}_{12}\ \cdots\ 2\bar{H}_{1\ell}\ \bar{H}_{22}\ \cdots\ 2\bar{H}_{2\ell}\ \cdots\ \bar{H}_{\ell\ell}\right]^T,$$
$$\bar{\xi}_k = \left[(\xi_k^1)^2\ \xi_k^1\xi_k^2\ \cdots\ \xi_k^1\xi_k^\ell\ (\xi_k^2)^2\ \cdots\ \xi_k^2\xi_k^\ell\ \cdots\ (\xi_k^\ell)^2\right]^T.$$

It should be emphasized that the unknown matrix $\bar{H} \in \mathbb{R}^{\ell\times\ell}$ has $\frac{1}{2}\ell(\ell+1)$ unknown elements due to $\bar{H}_{ij} = \bar{H}_{ji}$. From (40) and (41), one can obtain that

$$h^T \bar{\xi}_k = \bar{x}_k^T Q_x \bar{x}_k + u_k^T R u_k - a_k^T(\vartheta \otimes I)a_k + h^T \bar{\xi}_{k+1} \tag{42}$$

which is a key equation in the following Q-learning algorithms. Now, we are ready to use the Q-learning approach to learn the corresponding Q-function matrix $\bar{H}$. Policy iteration and value iteration algorithms using Q-learning technique are given in Algorithms 1 and 2, respectively. In Algorithms 1 and 2, a_k^j, $\phi_{k-1,k-N}^{j+1}$ and $\psi_{k-1,k-N}^{j+1}$ are defined as follows

$$\begin{aligned}a_k^j =& [(a_k^{1,j})^T, (a_k^{2,j})^T, \cdots, (a_k^{q,j})^T]^T \\ \phi_{k-1,k-N}^{j+1} =& \bar{H}_{a\bar{e}}^{j+1}\bar{e}_{k-1,k-N} + \bar{H}_{a\bar{u}}^{j+1}\bar{u}_{k-1,k-N} \\ &+ \bar{H}_{a\bar{a}}^{j+1}\bar{a}_{k-1,k-N} \\ \psi_{k-1,k-N}^{j+1} =& \bar{H}_{u\bar{e}}^{j+1}\bar{e}_{k-1,k-N} + \bar{H}_{u\bar{u}}^{j+1}\bar{u}_{k-1,k-N} \\ &+ \bar{H}_{u\bar{a}}^{j+1}\bar{a}_{k-1,k-N}\end{aligned}$$

$$Q(\xi_k)=\xi_k^T\begin{bmatrix} M_e^T H_{\bar{x}\bar{x}} M_e & M_e^T H_{\bar{x}\bar{x}} M_u & M_e^T H_{\bar{x}\bar{x}} M_a & M_e^T H_{\bar{x}u} & M_e^T H_{\bar{x}a} \\ M_u^T H_{\bar{x}\bar{x}} M_e & M_u^T H_{\bar{x}\bar{x}} M_u & M_u^T H_{\bar{x}\bar{x}} M_a & M_u^T H_{\bar{x}u} & M_u^T H_{\bar{x}a} \\ M_a^T H_{\bar{x}\bar{x}} M_e & M_a^T H_{\bar{x}\bar{x}} M_u & M_a^T H_{\bar{x}\bar{x}} M_a & M_a^T H_{\bar{x}u} & M_a^T H_{\bar{x}a} \\ H_{u\bar{x}} M_e & H_{u\bar{x}} M_u & H_{u\bar{x}} M_a & H_{uu} & H_{ua} \\ H_{a\bar{x}} M_e & H_{a\bar{x}} M_u & H_{a\bar{x}} M_a & H_{au} & H_{aa} \end{bmatrix}\xi_k=\xi_k^T\begin{bmatrix} \bar{H}_{\bar{e}\bar{e}} & \bar{H}_{\bar{e}\bar{u}} & \bar{H}_{\bar{e}\bar{a}} & \bar{H}_{\bar{e}u} & \bar{H}_{\bar{e}a} \\ \bar{H}_{\bar{u}\bar{e}} & \bar{H}_{\bar{u}\bar{u}} & \bar{H}_{\bar{u}\bar{a}} & \bar{H}_{\bar{u}u} & \bar{H}_{\bar{u}a} \\ \bar{H}_{\bar{a}\bar{e}} & \bar{H}_{\bar{a}\bar{u}} & \bar{H}_{\bar{a}\bar{a}} & \bar{H}_{\bar{a}u} & \bar{H}_{\bar{a}a} \\ \bar{H}_{u\bar{e}} & \bar{H}_{u\bar{u}} & \bar{H}_{u\bar{a}} & \bar{H}_{uu} & \bar{H}_{ua} \\ \bar{H}_{a\bar{e}} & \bar{H}_{a\bar{u}} & \bar{H}_{a\bar{a}} & \bar{H}_{au} & \bar{H}_{aa} \end{bmatrix}\xi_k \tag{37}$$

Algorithm 1 : Policy Iteration Using Q-Learning

1. **Initialization** : Set $j=0$, select stabilizing defender's policy u_k^0 and attackers' policy a_k^0, and choose $H^0=(H^0)^T$
2. **Policy Evaluation** : Solve for h^{j+1}
$-(h^{j+1})^T\bar{\xi}_k+\bar{x}_k^T Q_x\bar{x}_k+(u_k^j)^T R(u_k^j)$
$-(a_k^j)^T(\vartheta\otimes I)(a_k^j)+\gamma(h^{j+1})^T\bar{\xi}_{k+1}=0$
3. **Policy Improvement** :
$u_k^{j+1}=\left(\bar{H}_{uu}^{j+1}-\bar{H}_{ua}^{j+1}(\bar{H}_{aa}^{j+1})^{-1}\bar{H}_{au}^{j+1}\right)^{-1}$
$\times\left(\bar{H}_{ua}^{j+1}(\bar{H}_{aa}^{j+1})^{-1}\phi_{k-1,k-N}^{j+1}-\psi_{k-1,k-N}^{j+1}\right)$
$a_k^{j+1}=\left(\bar{H}_{aa}^{j+1}-\bar{H}_{au}^{j+1}(\bar{H}_{uu}^{j+1})^{-1}\bar{H}_{ua}^{j+1}\right)^{-1}$
$\times\left(\bar{H}_{au}^{j+1}(\bar{H}_{uu}^{j+1})^{-1}\psi_{k-1,k-N}^{j+1}-\phi_{k-1,k-N}^{j+1}\right)$
4. **Stop if**
$\|\bar{H}^{j+1}-\bar{H}^j\|<\epsilon$

Algorithm 2 : Value Iteration Using Q-Learning

1. **Initialization** : Set $j=0$, select any defender's policy u_k^0 and attackers' policy a_k^0, and choose $H^0=(H^0)^T$
2. **Policy Evaluation** : Solve for h^{j+1}
$(h^{j+1})^T\bar{\xi}_k=\bar{x}_k^T Q_x\bar{x}_k+(u_k^j)^T R(u_k^j)$
$-(a_k^j)^T(\vartheta\otimes I)(a_k^j)+\gamma(h^j)^T\bar{\xi}_{k+1}$
3. **Policy Improvement** :
$u_k^{j+1}=\left(\bar{H}_{uu}^{j+1}-\bar{H}_{ua}^{j+1}(\bar{H}_{aa}^{j+1})^{-1}\bar{H}_{au}^{j+1}\right)^{-1}$
$\times\left(\bar{H}_{ua}^{j+1}(\bar{H}_{aa}^{j+1})^{-1}\phi_{k-1,k-N}^{j+1}-\psi_{k-1,k-N}^{j+1}\right)$
$a_k^{j+1}=\left(\bar{H}_{aa}^{j+1}-\bar{H}_{au}^{j+1}(\bar{H}_{uu}^{j+1})^{-1}\bar{H}_{ua}^{j+1}\right)^{-1}$
$\times\left(\bar{H}_{au}^{j+1}(\bar{H}_{uu}^{j+1})^{-1}\psi_{k-1,k-N}^{j+1}-\phi_{k-1,k-N}^{j+1}\right)$
4. **Stop if**
$\|\bar{H}^{j+1}-\bar{H}^j\|<\epsilon$

In order to solve $\bar{H}^{j+1}$ by using (40) recursively, the number of samples $\bar{\xi}_k$ should satisfy $\upsilon\geq\frac{1}{2}\ell(\ell+1)$. Let $\Xi=\left[\bar{\xi}_k^1,\bar{\xi}_k^2,\cdots,\bar{\xi}_k^\upsilon\right]$ and $\Lambda=[\rho_k^1,\rho_k^2,\cdots,\rho_k^\upsilon]$, where $\rho_k^i=[\bar{x}_k^T Q_x\bar{x}_k+(u_k^j)^T R(u_k^j)-\vartheta(a_k^j)^T(a_k^j)]^i+\gamma(h^j)^T\bar{\xi}_{k+1}^i$, $i=1,2,\ldots,\upsilon$. Then, one can obtain that $(h^{j+1})^T\Xi=\Lambda$, which results in

$$h^{j+1}=\left(\Xi\Xi^T\right)^{-1}\Xi\Lambda^T \tag{43}$$

It is noted that, however, the defender's policy u_k and the attacker's policy a_k are dependent on the measured input-output $e_{k-1,k-N}$, $u_{k-1,k-N}$ and $a_{k-1,k-N}$. Consequently, the matrix $\Xi\Xi^T$ is not invertible. To address this issue, probing noise should be added into the system dynamics, and the following persistence of excitation (PE) condition must be satisfied.

Definition 4. A q-vector sequence $h=[h_1,h_2,\cdots,h_q]^T$ is said to be persistently exciting over an interval $[k+1,k+l]$ if for some constant $\delta>0$

$$\sum_{i=k+1}^{k+l}h_i(h_i)^T\geq\delta I \tag{44}$$

In order to satisfy the persistence of excitation condition, the actual defender's policy and attacker's policy are generated by

$$\hat{u}_k=u_k+n_k^1,\quad \hat{a}_k^j=a_k^j+n_k^{2,j} \tag{45}$$

where n_k^1 and $n_k^{2,j}$, $j=1,2,\ldots,q$, are probing noise signals.

Remark 5. If the state almost converges to the desired position and becomes stationary, then the persistence of excitation condition is no longer satisfied. An exploratory signal consisting of sinusoids of varying frequencies can be added to the policies of the defender and attackers to ensure PE qualitatively. Consequently, n_k^1 and $n_k^{2,j}$ in (45) can be chosen as sinusoids of varying frequencies, exponentially decaying noise, or Gaussian noise.

According to the method given in Rizvi et al. (2018), the following theorem can be easily obtained, and the corresponding proof is omitted here.

Theorem 6. Assume the system is controllable and observable. The sequences u_k^j and $a_k^{i,j}$, $i\in\mathcal{Q}$, $j=1,2,\ldots,\infty$, generated by Q-learning Algorithm 7 or Algorithm 8, can converge to the optimal strategies (38) and (39) for the defender and the attackers when the system is sufficiently excited.

4. AN ILLUSTRATIVE EXAMPLE

In this section, an example is given to demonstrate the main results proposed in this paper.

A linear discrete-time system is given as

$$x_{k+1}=Ax_k+Bu_k+B(\Gamma^1a_k^1+\Gamma^2a_k^2)$$
$$y_k=Cx_k$$

with

$$A=\begin{bmatrix}-1.5 & 0.3 & 1\\ 1.1 & 0.7 & -0.5\\ 0.5 & -0.2 & 1.9\end{bmatrix},\ B=\begin{bmatrix}0.2 & -0.3\\ 0.6 & -1\\ -1.1 & 0.8\end{bmatrix},$$
$$\Gamma^1=\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix},\ \Gamma^2=\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix},\ C=[1\quad 1\quad 1].$$

Note that the eigenvalues of system matrix A are $\lambda_1=-1.8021$, $\lambda_2=0.8380$ and $\lambda_3=2.0641$. Therefore, the open-loop system is unstable. The reference trajectory is generated by

$$y_{k+1}^r=-y_k^r$$

Other parameters are given as follow.

$$\gamma=0.88,\ \theta_1=0.83,\ \theta_2=0.81,\ \epsilon=10^{-6},$$
$$Q_e=0.9,\ R=\begin{bmatrix}0.25 & 0\\ 0 & 0.001\end{bmatrix}.$$

The simulation results can be obtained according to Algorithm 2 and are depicted in Fig.1 - Fig.4, which show that the tracking error can converge to zero, i.e., the system output y_k can track the reference input y_k^r.

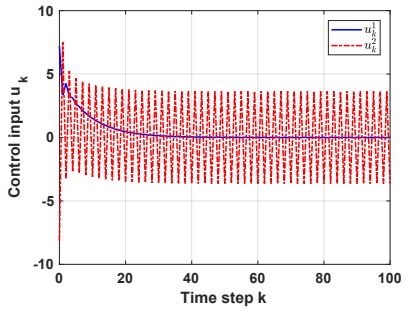

Fig. 1. Control input u_k

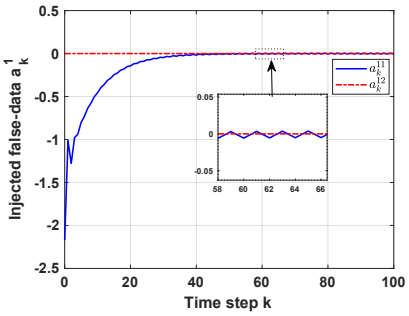

Fig. 2. False-data injected by attacker 1

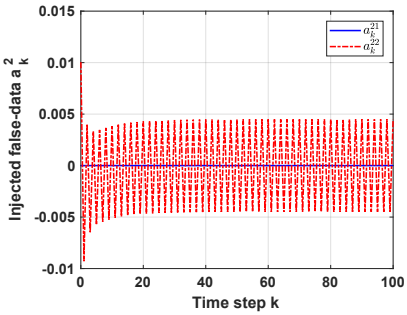

Fig. 3. False-data injected by attacker 2

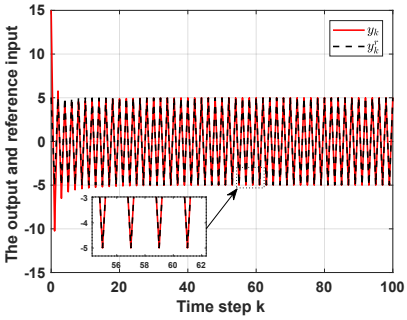

Fig. 4. System output and the reference trajectory

5. CONCLUSION

In this paper, game theory was used to investigated the optimal tracking control problem in the presence of false-data-injection attacks. Then, the Q-learning method was developed to solve the GARE online without requiring the knowledge of augmented system dynamics. Moreover, the Q-function was expressed in terms of only measured input-output data, and the policies for both sides were generated by Q-learning algorithm, where the system was assumed to be sufficiently excited. The simulation results have shown that the system output can track the given reference trajectory under FDI attacks.

REFERENCES

A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson. A secure control framework for resource-limited adversaries, *Automatica*, 51, 135–148, 2015.

D. Persis, and P. Tesi. Networked control of nonlinear systems under denial-of-service, *Syst. Contr. Lett.*, 96, 124–131, 2016.

J. H. Qin, M. L. Li, L. Shi, and X. H. Yu. Optimal denial-of-service attack scheduling with energy constraint over packet-dropping networks, *IEEE Trans. Autom. Control*, 63(6), 1648–1663, 2018.

K. M. Ding, Y. Z. Li, D. E. Quevedo, S. Dey, and L. Shi. A multi-channel transmission schedule for remote state estimation under DoS attacks, *Automatica*, 78, 194–201, 2017.

E. Kung, S. Dey, and L. Shi. The performance and limitations of ϵ-stealthy attacks on higher order systems, *IEEE Trans. Autom. Control*, 62(2), 941–947, 2017.

C. Z. Bai, V. Gupta, and F. Pasqualetti. On Kalman filtering with compromised sensors: Attacks stealthiness and performance bounds, *IEEE Trans. Autom. Control*, 62(12), 6641–6648, 2017.

L. Hu, Z. D. Wang, Q. L. Han, and X. H. Liu. State estimation under false data injection attacks: Security analysis and system protection, *Automatica*, 87, 176–183, 2018.

Y. L. Mo, and B. Sinopoli. Secure control against replay attacks, *In Proc. 47 Annual Allerton Conference*, pp. 911-918, 2009.

M. H. Zhu, and S. Martnez. On the performance analysis of resilient networked control systems under replay attacks, *IEEE Trans. Autom. Control*, 59(3), 804–808, 2014.

B. Chen, D. W. C. Ho, G. Hu, and L. Yu. Secure fusion estimation for bandwidth constrained cyber-physical systems under replay attacks, *IEEE Trans. Cybern.*, 48(6), 1862–1876, 2018.

C. Z. Bai, F. Pasqualetti, and V. Gupta. Data-injection attacks in stochastic control systems: Detectability and performance tradeoffs, *Automatica*, 82, 251–260, 2017.

Y. Z. Li, D. W. Shi, and T. W. Chen. False data injection attacks on networked control systems: A Stackelberg-game analysis, *IEEE Trans. Autom. Control*, 63(10), 3503–3509, 2018.

S. A. A. Rizvi, and Z. L. Lin. Output feedback Q-learning for discrete-time linear zero-sum games with application to the H-infinity control, *Automatica*, 95, 213–221, 2018.

D. Ye, S. P. Luo. A co-design methodology for cyber-physical systems under actuator fault and cyber attack, *J. Franklin Inst.*, 356, 1856–1879, 2019.

D. R. Ding, Z. D. Wang, D. W. C. Ho, and G. L Wei. Observer-based event triggering consensus control for multi-agent systems with lossy sensors and cyber attacks, *IEEE Trans. Cybern.*, 47(8), 1936–1947, 2017.