Discrimination between Replay Attacks and Sensor Faults for Cyber-Physical Systems via Event-triggered Communication[★]

Kangkang Zhang[a,b,*], Christodoulos Keliris[a,b], Marios M. Polycarpou[a,b] and Thomas Parisini[a,c,d]

[a] KIOS Research and Innovation Center of Excellence
[b] Dept. of Electrical and Computer Engineering, University of Cyprus, Nicosia, 1678, Cyprus
[c] Dept. of Electrical and Electronic Engineering, Imperial College London, London, SW7 2AZ, UK
[d] Dept. of Engineering and Architecture, University of Trieste, Trieste, 34127, Italy

ARTICLE INFO

Keywords:
Threat discrimination
replay attacks
sensor bias faults
event-triggered communication

ABSTRACT

In this paper, a threat discrimination methodology is proposed for cyber-physical systems with event-triggered data communication, aiming to identify sensor bias faults from two possible types of threats: replay attacks and sensor bias faults. Event-triggered adaptive estimation and backward-in-time signal processing are the main techniques used. Specifically, distinct incremental systems of the event-triggered cyber-physical system resulting from the considered threat types are established for each threat type, and the difference between their inputs are found and utilized to discriminate the threats. An event-triggered adaptive estimator is then designed by using the event-triggered sampled data based on the system in the attack case, allowing to reconstruct the unknown increments in both the threat cases. The backward-in-time model of the incremental system in the replay attack case is proposed as the signal processor to process the reconstructions of the increments. Such a model can utilize the aforementioned input difference between the incremental systems such that its output has distinct quantitative properties in the attack case and in the fault case. The fault discrimination condition is rigorously investigated and characterizes quantitatively the class of distinguishable sensor bias faults. Finally, a numerical simulation is presented to illustrate the effectiveness of the proposed methodology.

1. Introduction

Cyber-physical systems (CPS) have attracted many research efforts recently owing to their wide applications. However, CPS are highly vulnerable to malicious cyber attacks as a result of integrating computation, communication and control [3]. Therefore, state-of-the-art cyber attack diagnosis technologies should be developed.

In the past decade, model-based detection of integrity attacks such as covert attacks [21, 1], zero-dynamics attacks [25] and replay attacks [19], has been investigated by the research community [27]. Several survey papers, such as Dibaji et al. [5] and Ding et al. [6], have detailed the main model-based detection methodologies. Replay attacks are commonly used by malicious adversaries due to their simplicity in implementation. The famous "Stuxnet worm" virus occurring in the Iranian nuclear facilities is a typical implementation example of replay attacks [11]. In a replay attack event, the attacker records the data from the normal plant operation over a time interval and then replays the data to the supervisory system [22]. Hence, replay attacks possess high stealthiness properties with respect to most of the traditional anomaly detectors such as the ones in Ding [7] and Blanke et al. [2] as a result of the used malicious data taken from the normal system operation. In general, replay attacks are usually performed along with other types of un-stealthy cyber attacks, such as data-injection attacks, aiming to cover them up. In more detail, during the replaying procedure of a replay attack, the other un-stealthy attacks can deteriorate significantly the operation of the system, at the same time, remain concealed from the typical anomaly detectors due to the cover provided by the replay attack. Therefore, the main objective of replay attacks is to hide the additional attack events that are launched during the replaying procedure of the replay attacks. Watermarking is the main methodology to detect replay attacks, which is proposed in [19] and is achieved by adding watermarks to the control inputs. However, such additive watermarks may cause control performance degradation. A model inversion-based watermarking is proposed in [20] to alleviate this issue. Furthermore, a multiplicative watermarking approach is proposed in [9] and recently extended in [10].

The threat discrimination (TD) problem arises when taking into account two types of threats: physical faults and cyber attacks. The aim of TD is to identify the occurring threat type, namely determining which type of threats (attacks or faults) is occurring. However, in the aforementioned literature and references therein, the TD problem is rarely mentioned and is generally overseen by the research communities. TD is very important for practical applications, since it can help operators make correct decisions and take suit-

[★] This work has been supported by: European Union's Horizon 2020 Research and Innovation Program under the grant agreement no. 739551 (KIOS CoE), the Italian Ministry for Research in the framework of the 2017 Program for Research Projects of National Interest (PRIN) (grant no. 2017YKXYXJ), the National Natural Science Foundation of China (grant no. 61903188), the Natural Science Foundation of Jiangsu Province (grant no. BK20190403) and the China Postdoctoral Science Foundation (grant no. 2019M650114).

[*] Corresponding author
✉ kzhang02@ucy.ac.cy (K. Zhang); keliris.chris@gmail.com (C. Keliris); mpolycar@ucy.ac.cy (M.M. Polycarpou); t.parisini@gmail.com (T. Parisini)
ORCID(s):

able remediation actions against threats. Mitigation strategies against cyber attacks and physical faults are usually different. Physical maintenance, such as replacing communication cables, can be effective in mitigating physical faults, but cannot remediate the issues caused by cyber attacks. Updating communication protocols and firewalls are the general prevention approaches against cyber attacks. In the case of a cyber attack, it may be preferable to take drastic actions such as shutting down the whole system rather than trying to fix it alongside the actions of the attacker. Therefore, TD identifies the type of the occurring threats, thereby guiding the corresponding remediation measures for avoiding catastrophic consequences. The TD problem was first considered in Chanthery and Subias [4], in which simulation results of distinguishing between faults and attacks in a two-tank benchmark based on some traditional model-based and data-based anomaly diagnosis methods were presented. A TD scheme is also designed in Taheri et al. [24] by applying filters in both the plant side and the control side of the closed-loop CPS. Such a TD scheme may increase the communication load because the filter signals in the plant side should be independently transmitted to the control side. Typical fault detection and isolation methods cannot be exploited directly to solve the TD problem, which is also the reason that the TD problem between replay attacks and physical sensor bias faults remains an open problem. One challenge of TD is the stealthiness of replay attacks. In general, typical fault detection and isolation methods do not consider the stealthiness issue by design, which prevents them from sensing the stealthy replay attacks. Another challenge of TD is that the time responses of a detector to replay attacks possess distinct characteristics with respect to the responses of the same detector to physical sensor faults. Traditional fault detection and isolation methods may not consider sufficiently such time response characteristics and thus, cannot take full advantage of the difference between replay attacks and sensor bias faults. Therefore, based on the aforementioned challenges, traditional anomaly detection and isolation methods may be not able to discriminate between replay attacks and sensor bias faults.

Event-triggering techniques are proposed to save communication resources and are generally used in the fields of control and observer design for complex systems, such as [23, 18, 13]). Model-based fault diagnosis problem for event-triggered control system has also been studied in recent years. Many related results, such as [29, 16], have dealt with this problem. However, only a few results take into account both the attack diagnosis and the event-triggered communication. A sliding mode observer based attack detection and estimation scheme is proposed in [14] for linear autonomous vehicle platoons using the event-triggered communication.

This paper develops an attack and fault discrimination methodology using event-triggered communication data to distinguish between replay attacks and physical sensor faults. Specifically, a threat discrimination methodology is proposed in this paper using event-triggered communication data for identifying the case of sensor bias faults between the two types of threats: replay attacks and sensor bias faults. An event-triggered adaptive estimator and a backward-in-time model are designed to achieve the discrimination task. In particularly, the incremental systems of the event-triggered CPS resulting from the replay attacks and the sensor faults are established and their different inputs are found. An event-triggered adaptive estimator is designed using the event-triggered communication data based on the structural characteristics of the CPS in the attack case, allowing to reconstruct the unknown increments in both of the threat cases. The backward-in-time model of the incremental system in the attack case is then proposed as the signal processor to process the reconstruction of the increments, which can utilize the distinct inputs of the incremental systems to generate distinguishable outputs. Such a backward-in-time model can guarantee that in the presence of the replay attacks, its output is lower than a threshold, whereas in the fault case its output exceeds the threshold, and therefore, allowing to identify the sensor bias faults. The fault discrimination condition is rigorously investigated and characterizes quantitatively the class of distinguishable faults. In conclusion, the main contributions of this paper are summarized as follows:

- A threat discrimination framework consisting of an estimator and a backward-in-time signal processor is proposed and realized, for identifying sensor bias faults between two threat scenarios, namely replay attacks and sensor bias faults;
- Using the event-triggered communication data, an event-triggered adaptive estimator is designed to provide the estimates of the system states, output transmission errors and the threat parameters for reconstructing the unknown system increments;
- The backward-in-time model of the incremental system due to the replay attack is introduced, which is able to utilize the distinct increments due to the replay attack and the sensor bias faults and generate distinguishable outputs.

The rest sections of this paper are organized as follows. In Section 2, the problem formulation is given. In Section 3, the threat discrimination scheme including an event-triggered estimator and a backward-in-time model is designed and rigorously analyzed, and a numerical simulation is presented in Section 4. Finally, the conclusions are drawn in Section 5.
Notations: The notation $|\cdot|$ is used in this paper to represent the absolute value for scalars, and the 2-norm for vectors and matrices. For a square matrix $A \in \mathbb{R}^{n\times n}$, $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ represent the maximum eigenvalue of the minimum eigenvalue of A respectively.

2. Problem Formulation

2.1. CPS with Event-Triggered Communication

In this paper, we consider a type of CPS depicted in Fig. 1, which consists of a physical plant $\mathcal{P}$, an event-triggering

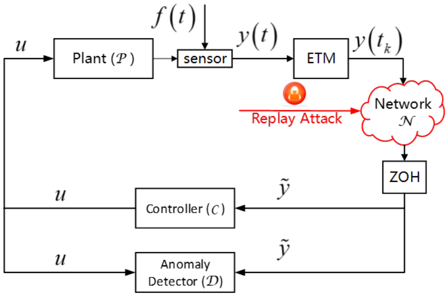

Figure 1: Structure diagram of CPS with event-triggered communication in the presence of replay attacks.

mechanism (ETM), a data transmission network $\mathcal{N}$, a zero-order holder (ZOH), a controller $\mathcal{C}$ and an anomaly detector $\mathcal{D}$. The physical plant $\mathcal{P}$ is a linear time-invariant system, and $\mathcal{C}$ is a linear static output-feedback controller. The plant $\mathcal{P}$ and $\mathcal{C}$ in the nominal case (no attack or fault) are described by

$$\begin{aligned} \dot{x}_n(t) &= Ax_n(t) + BK\tilde{y}_n(t), &(1)\\ y_n(t) &= Cx_n(t), &(2) \end{aligned}$$

where $x_n \in \mathbb{R}^{n_p}$ is the state of the physical plant in the nominal case, $y_n \in \mathbb{R}^{n_y}$ represents the sensor measurements and $\tilde{y}_n \in \mathbb{R}^{n_y}$ is the sampled sensor measurements. The matrices $A \in \mathbb{R}^{n_p \times n_p}$, $B \in \mathbb{R}^{n_p \times n_u}$, $K \in \mathbb{R}^{n_u \times n_y}$ and $C \in \mathbb{R}^{n_y \times n_p}$ are known by the defender. We suppose that A is a Hurwitz matrix and the pair (A, C) is observable. In the case that A is not a Hurwitz matrix, the discrimination schemes in this paper can be developed based on an asymptotically stable observer, which, however, will lead to complicated notations and presentation. Hence, for simplification purposes, the matrix A is assumed to be a Hurwitz matrix throughout the paper.

For simplicity, only the sensor data communication network $\mathcal{N}$ and the ETM in the sensor-to-controller channel are considered. The task of the ETM is to determine whether the data at the current time instant is transmitted or not, thereby scheduling the transmission of $\mathcal{N}$ and reducing the load. The event-triggering condition of the ETM in the nominal case is given by

$$t^n_{k+1} = \inf\left\{t > t^n_k \,|\, |y_n(t^n_k) - y_n(t)|^2 > \delta |y_n(t)|^2\right\}, \quad (3)$$

where $\delta > 0$ is a pre-specified threshold. It should be noted that the event-triggering instants $t^n_1, t^n_2, \cdots, t^n_k$ can be exactly indicated by the jumps of the signal $\tilde{y}_n(t)$. Thus, $t_1, t_2, \cdots, t_k$ can be obtained in real time.

Remark 1. *In general, the event-triggering condition is evaluated using the actual output measurement $y(t_k)$ and $y(t)$ rather than the corresponding nominal values indicated in (3), since $y_n(t^n_k)$ and $y_n(t)$ in the presence of a threat are unknown.* ▼

Based on the transmission time instants of the ETM, the ZOH possesses a holding time $t \in \Gamma^n_k \triangleq [t^n_k, t^n_{k+1})$, and retains its input during Γ^n_k. Note that y_n is the input of the ETM, and $\tilde{y}_n$ is the output of the ZOH. The transmission of the ETM and ZOH can be characterized by $\tilde{y}_n(t) = y_n(t^n_k) = Cx(t^n_k)$ for $t \in \Gamma^n_k$. Thus, the transmission error η_n of the ETM and ZOH can be defined by

$$\begin{aligned} \eta_n(t) &\triangleq \tilde{y}_n(t) - y_n(t) \\ &= Cx_n(t^n_k) - Cx_n(t), \ \forall\, t \in \Gamma^n_k, &(4)\\ \eta^+_n(t^n_{k+1}) &= 0, \ \forall\, t = t^n_k, \ k \in \mathbb{Z}_+, &(5) \end{aligned}$$

where $\eta^+_n(t)$ is given in the footnote[1], and with the defined η_n, $\tilde{y}_n$ between two consecutive transmission instants can be described by

$$\tilde{y}_n(t) = y_n(t^n_k) = y_n(t) + \eta_n(t), \ \forall\, t \in \Gamma^n_k. \quad (6)$$

Therefore, by synthesizing (1)-(6), the closed-loop CPS in the nominal case is described by the following impulsive system:

$$\mathcal{W}_n : \left\{ \begin{array}{rcl} \dot{x}_n(t) &=& Ax_n(t) + BK\tilde{y}_n(t), \\ \dot{\eta}_n(t) &=& -CAx_n(t) - CBK\tilde{y}_n(t), \\ \eta^+_n(t^n_k) &=& 0, \\ y_n(t) &=& Cx_n(t), \\ \tilde{y}_n(t) &=& y_n(t) + \eta_n(t). \end{array} \right. \quad (7)$$

2.2. Potential Threats

In this paper, we consider two types of threats: 1) replay attack in the communication network $\mathcal{N}$ and 2) single or multiple sensor bias fault(s). The modeling of such threats will be described in the sequel. To this end, in order to distinguish the variables in the attack case and in the fault case, the subscripts a and f are used respectively. For example, x_a and x_f represent the plant states in the attack case and in the fault case respectively.

2.2.1. Replay Attack Scenarios.

In general, replay attacks have recording and replaying procedures. The adversary first records the sensor measurement y_n communicated through $\mathcal{N}$ starting at a time $T_a - T$ and for a recording time T, so that the recording ends at the time T_a. Then, the replaying procedure starts from T_a and lasts till $T_a + T$. In this paper, for the sake of simplifying the presentation, we only consider a single replay and no repeat. In addition, we consider that T is sufficiently long such that the developed scheme has enough time to carry out the discrimination task. Therefore, the replaying procedure is limited in the time interval $[T_a,\ T_a + T)$, and the network $\mathcal{N}$ recovers to the normal operation after $T_a + T$.

Let Γ_a represent the time interval of the replaying procedure, i.e., $\Gamma_a \triangleq [T_a,\ T_a + T)$. For a time-varying variable $p(t)$, we use $p'(t)$ to represent its value at the time $t - T$, i.e., $p'(t) \triangleq p(t - T)$. Then, in the presence of the replay attack, we have

$$\tilde{y}_a(t) = \tilde{y}'_n(t), \ \forall\, t \in \Gamma_a. \quad (8)$$

[1] For a signal $x : [0, +\infty) \to \mathbb{R}^n$, we denote the limit from above at time $t \in [0, +\infty)$ by $x^+(t) = \lim_{s \downarrow t} x(s)$.

Moreover, the data transmission time instants specified by the event-triggering condition (3) may shift in the presence of the replay attack (see *Remark 1*), which are now denoted by t_k^a with $k \in \mathbb{Z}_+$, and the time interval between two consecutive transmission instants is denoted by $\Gamma_k^a \triangleq [t_k^a,\ t_{k+1}^a)$. Then, it follows from (6) and (8) that

$$\tilde{y}_a(t) = \tilde{y}_n'(t) = y_n'(t) + \eta_n'(t),\ \forall\, t \in \Gamma_k^a. \tag{9}$$

Now, a virtual attack signal $a(t)$ of the replay attack is defined as follows

$$Ca(t) \triangleq Cx_n'(t) + \eta_n'(t) - Cx_a(t) - \eta_a(t), \tag{10}$$

where x_a is the state of the plant in the attack case, and η_a is defined based on (4) as follows:

$$\eta_a(t) \triangleq Cx_a(t_k^a) - Cx_a(t),\ \forall\, t \in \Gamma_k^a. \tag{11}$$

Then, $\mathcal{W}_n$ in (7) in the attack case can be described by the following impulsive system:

$$\mathcal{W}_a : \left\{ \begin{array}{rcl} \dot{x}_a(t) & = & Ax_a(t) + BK\tilde{y}_a(t), \\ \dot{\eta}_a(t) & = & -CAx_a(t) - CBK\tilde{y}_a(t), \\ \eta_a^+(t_k^a) & = & 0, \\ y_a(t) & = & Cx_a(t), \\ \tilde{y}_a(t) & = & Cx_a(t) + \eta_a(t) + Ca(t). \end{array} \right. \tag{12}$$

In order to guarantee the stealthiness, $a(t)$ should be sufficiently small, and hence, the following assumption gives the boundedness restriction on $a(t)$.

Assumption 1. *The attack signal $a(t)$ is sufficiently small in amplitude, i.e.,*

$$a(t) \in \Theta^a \triangleq \left\{ \theta \in \mathbb{R}^{n_p} \,\middle|\, |\theta| \leq \sigma_a, \sigma_a > 0 \right\},\ \forall\, t \in \Gamma_a, \tag{13}$$

where σ_a is sufficiently small and known by the defender. ▼

2.2.2. Sensor Bias Fault Scenarios.

In this work, in order to simplify the presentation, we consider the case of a single/multiple sensor bias fault(s) occurring simultaneously at time T_f. Let Γ_f denote the time interval after the occurrence of the sensor faults, i.e., $\Gamma_f \triangleq [T_f, +\infty)$. Then, the sensor measurement is described by

$$y_f(t) = Cx_f(t) + Cf(t),\ \forall\, t \in \Gamma_f, \tag{14}$$

where x_f is the state of the plant in the presence of the fault and $f : [0, +\infty) \rightarrow \mathbb{R}^{n_p}$ represents the sensor bias fault function. Moreover, the time transmission instants specified by the event-triggering condition (3) may also shift in the presence of the sensor bias faults (see *Remark 1*). The new time transmission instants in the fault case are denoted by t_k^f with $k \in \mathbb{Z}_+$, and the time interval between two consecutive transmission instants is denoted by $\Gamma_k^f \triangleq [t_k^f,\ t_{k+1}^f)$. By defining η_f as

$$\eta_f(t) \triangleq Cx_f(t_k^f) - Cx_f(t),\ \forall\, t \in \Gamma_k^f, \tag{15}$$

it follows from (6) and (14) that

$$\tilde{y}_f(t) = y_f(t_k^f) = Cx_f(t) + \eta_f(t) + Cf(t_k^f),\ \forall\, t \in \Gamma_k^f. \tag{16}$$

Consequently, in the presence of the sensor bias fault, $\mathcal{W}_n$ in (7) can be described by the following impulsive system:

$$\mathcal{W}_f : \left\{ \begin{array}{rcl} \dot{x}_f(t) & = & Ax_f(t) + BK\tilde{y}_f(t), \\ \dot{\eta}_f(t) & = & -CAx_f(t) - CBK\tilde{y}_f(t), \\ \eta_f^+(t_k^f) & = & 0, \\ y_f(t) & = & Cx_f(t) + Cf(t), \\ \tilde{y}_f(t) & = & Cx_f(t) + \eta_f(t) + Cf(t_k^f). \end{array} \right. \tag{17}$$

In addition, the sensor bias fault vector $f(t)$ is assumed to satisfy the following assumption.

Assumption 2. *The fault vector $f(t)$ is norm bounded, i.e.,*

$$f(t) \in \Theta^f \triangleq \left\{ \theta \in \mathbb{R}^{n_p} \,\middle|\, |\theta| \leq \sigma_f, \sigma_f > 0 \right\},\ \forall\, t \in \Gamma_f, \tag{18}$$

where σ_f is known by the defender. ▼

Remark 2. *The boundedness requirement for $f(t)$ in Assumption 2 is commonly used in the fault diagnosis literature (see, e.g., [28]).* ∇

Regarding the threat scenarios considered in this paper, we have the following assumption.

Assumption 3. *It is assumed that only one type of threat can occur in the system: either a replay attack or sensor bias fault(s).* ▼

2.3. Objective

Suppose that a threat has been detected by the anomaly detector $\mathcal{D}$ in Fig. 1 at the time T_d where $T_d \geq T_a$ and $T_d \geq T_f$, but the type of such a threat can not be identified by $\mathcal{D}$. The aim of this paper is to design a scheme to identify the sensor bias fault(s), i.e., whether the occurring threat is the sensor bias fault(s).

3. Threat discrimination Scheme

In this section, a framework is proposed to distinguish the sensor bias faults from the two considered threat scenarios. Based on this framework, the discrimination scheme consisting of an estimator and a signal processing model are designed and rigorously investigated.

3.1. Threat Discrimination Framework

3.1.1. Incremental Systems.

Considering the replay attack case, the virtual attack model is first presented in the sequel. The attack signal during the replaying procedure is the sensor measurements y_n of $\mathcal{W}_n$ in (7) during the recording procedure. Thus, the virtual attack model implemented by the attacker is the nominal system $\mathcal{W}_n$ during the recording procedure [19], i.e.,

$$\mathcal{W}'_n : \begin{cases} \dot{x}'_n(t) &= Ax'_n(t) + BK\tilde{y}'_n(t), \\ y'_n(t) &= Cx'_n(t), \\ \tilde{y}'_n(t) &= y'_n(t) + \eta'_n(t), \end{cases} \tag{19}$$

where, as mentioned previously, $p'(t)$ is defined as $p'(t) = p(t - T)$ for $p \in \{x_n, \tilde{y}_n, y_n, \eta_n\}$.

The incremental system due to the replay attack is defined as the deviation between $\mathcal{W}_a$ and $\mathcal{W}'_n$. In particular, for a time-varying variable $q(t)$ ($q \in \{x, y, e, \tilde{y}\}$), the increment is defined as

$$\Delta q_a(t) \triangleq q_a(t) - q'_n(t). \tag{20}$$

According to this definition, we have

$$\Delta x_a = x_a - x'_n,\ \Delta y_a = C\Delta x_a,\ \Delta\eta_a = \eta_a - \eta'_n,\ \Delta\tilde{y}_a = 0.$$

Then, from $\mathcal{W}_a$ in (12) and $\mathcal{W}'_n$ in (19), the incremental system can be obtained as

$$\Delta\mathcal{W}_a : \begin{cases} \Delta\dot{x}_a(t) &= A\Delta x_a(t), \\ \Delta y_a(t) &= C\Delta x_a(t), \\ \Delta\tilde{y}_a(t) &= 0. \end{cases} \tag{21}$$

We now proceed to define the incremental system in the sensor fault case. The incremental system is defined as the deviation between $\mathcal{W}_f$ and $\mathcal{W}_n$. Specifically, for a time-varying variable $q(t)$ ($q \in \{x, y, \eta, \tilde{y}\}$), the change is defined as

$$\Delta q_f(t) \triangleq q_f(t) - q_n(t). \tag{22}$$

According to this definition, we have

$$\Delta x_f = x_f - x_n,\ \Delta y_f = C\Delta x_f + Cf,$$
$$\Delta\eta_f = \eta_f - \eta_n,\ \Delta\tilde{y}_f = C\Delta x_f + \Delta\eta_f + Cf(t_k^f).$$

Hence, from $\mathcal{W}_f$ in (17) and $\mathcal{W}_n$ in (7), the incremental system due to the sensor bias fault can be obtained as

$$\Delta\mathcal{W}_f : \begin{cases} \Delta\dot{x}_f(t) = A\Delta x_f(t) + BK\Delta\tilde{y}_f(t), \\ \Delta y_f(t) = C\Delta x_f(t) + Cf(t), \\ \Delta\tilde{y}_f(t) = C\Delta x_f(t) + \Delta\eta_f(t) + Cf(t_k^f). \end{cases} \tag{23}$$

The following proposition is derived based on the above definitions (20) and (22) of the increments.

Proposition 1. *(i) In the context of the definition (20), the virtual attack signal $a(t)$ satisfies*

$$a(t) = -\Delta x_a(t_k^a),\ \forall\, t \in \Gamma_k^a. \tag{24}$$

(ii) In the context of the definitions (20) and (22), $\Delta x_a(t)$ and $\Delta x_f(t)$ can be jointly described by

$$\Delta x_i(t) = e^{A(t-T_d)}\Delta x_i(T_d) + \int_{T_d}^{t} e^{A(t-\tau)} BK\Delta\tilde{y}_i(\tau)d\tau, \tag{25}$$

where $\Delta\tilde{y}_a = 0$ in (21) and $\Delta\tilde{y}_f$ is given in (23). ■

Proof. (i) According to the definition in (10), $Ca(t)$ in (10) can be equivalently written as

$$Ca(t) = -C\Delta x_a(t) - \Delta\eta_a(t) = -C\Delta x_a(t_k^a),\ \forall\, t \in \Gamma_k^a,$$

where $\Delta x_a(t) = x_a(t) - x'_n(t)$ and $\Delta\eta_a(t) = Cx_a(t_k^a) - Cx_a(t) - (Cx'_n(t_k^a) - Cx'_n(t)) = C(\Delta x_a(t_k^a) - \Delta x_a(t))$ are used. Thus, (24) follows.

(ii) According to (21), for $t > T_d$, $\Delta x_a(t)$ is described by

$$\Delta x_a(t) = e^{A(t-T_d)}\Delta x_a(T_d).$$

By solving the differential equation in (23), $\Delta x_f(t)$ can be described by (25). Hence, result (ii) follows. □

3.1.2. *Technical Framework.*

Considering the incremental systems $\Delta\mathcal{W}_a$ in (21) and $\Delta\mathcal{W}_f$ in (23), the difference between $\Delta\tilde{y}_a = 0$ in $\Delta\mathcal{W}_a$ and $\Delta\tilde{y}_f \neq 0$ in $\Delta\mathcal{W}_f$ is utilized in this paper to achieve the threat discrimination task. More specifically, the threat discrimination framework integrates an event-triggered adaptive estimator and a backward-in-time signal processor in a cascade way. The estimator is proposed based on the structure of $\mathcal{W}_a$ such that Δx_a can be reconstructed more accurately than Δx_f. The task of the backward-in-time signal processor is to process the reconstructions of Δx_i, $i = \{a, f\}$. This signal processor is designed as the backward-in-time model of $\Delta\mathcal{W}_a$ such that in the presence of Δx_a, its output is lower than a threshold, whereas under Δx_f the output potentially exceeds the threshold. Hence, the sensor bias fault can be identified.

3.2. Event-triggered Adaptive Estimator Design

In this subsection, an event-triggered adaptive estimator will be designed based on the structure of $\mathcal{W}_a$ in (12). To this end, a unified form of $\mathcal{W}_a$ in (12) and $\mathcal{W}_f$ in (17) are given. We start by writing $a(t)$ in (10) in a linear parameterization form. It follows from (25) in Proposition 1 that $\Delta x_a(t)$ can be written as $\Delta x_a(t) = e^{A(t-T_d)}\Delta x_a(T_d)$. By letting $\theta^a \triangleq \Delta x_a(T_d)$ and $F^a(t) \triangleq -e^{A(t-T_d)}$, then based on (24) in Proposition 1, we have

$$a(t) = F^a(t_k^a)\theta^a,\ \forall\, t \in \Gamma_k^a. \tag{26}$$

To maintain form consistence with $a(t)$, $f(t_k^f)$ in $\mathcal{W}_f$ is also written in the linear parameterization form as follows:

$$f(t_k^f) = [f_1(t_k^f), \cdots, f_{n_p}(t_k^f)]^T = F_f(t_k^f)\theta_f, \tag{27}$$

where $F_f(t_k^f) \triangleq \operatorname{diag}\{f_1(t_k^f), \cdots, f_{n_p}(t_k^f)\} \in \mathbb{R}^{n_p \times n_p}$ and $\theta^f \triangleq [1, 1, \cdots, 1]^T \in \mathbb{R}^{n_p}$. Therefore, $\mathcal{W}_a$ in (12) and $\mathcal{W}_f$ in (17) can be written in the following unified form:

$$\mathcal{W} : \begin{cases} \dot{x}(t) &= Ax(t) + BK\tilde{y}(t), \\ \dot{\eta}(t) &= -CAx(t) - CBK\tilde{y}(t), \\ \eta^+(t_k) &= 0, \\ \tilde{y}(t) &= C_0[x(t), \eta(t)] + CF(t_k)\theta, \end{cases} \tag{28}$$

where $C_0 \triangleq [C, I]$, x, η, $\tilde{y}$, t_k and $F(t_k)\theta$ are given in Tab. 1. Such a system $\mathcal{W}$ can represent $\mathcal{W}_a$ in the attack case and can also represent $\mathcal{W}_f$ in the fault case.

Table 1
Unified notations.

Notation	Fault case	Attack case
x	x_f	x_a
η	η_f	η_a
$\tilde{y}$	$\tilde{y}_f$	$\tilde{y}_a$
t_k	t_k^f	t_k^a
$F(t_k)\theta$	$F^f(t_k^f)\theta^f$	$F^a(t_k^a)\theta^a$

Next, the threat discrimination estimator activated at time T_d is designed based on $\mathcal{W}$ given in (28) in the attack case. Let $\hat{x}$, $\hat{\eta}$, $\hat{\tilde{y}}$ and $\hat{\theta}$ be the estimates of x, η, $\tilde{y}$ and θ, respectively. Then, the discrimination estimator is proposed as follows:

$$\mathcal{E}: \begin{cases} \dot{\hat{x}} &= A\hat{x} + L_x(\hat{\tilde{y}} - \tilde{y}) + BK\tilde{y} + \Omega_x\dot{\hat{\theta}}, \\ \dot{\hat{\eta}} &= -CA\hat{x} + L_e(\hat{\tilde{y}} - \tilde{y}) - CBK\tilde{y} + \Omega_\eta\dot{\hat{\theta}}, \\ \dot{\Omega}_x &= (A + L_xC)\Omega_x + L_x\Omega_\eta + L_xCF^a(t_k), \\ \dot{\Omega}_\eta &= (-CA + L_eC)\Omega_x + L_e\Omega_\eta + L_eCF^a(t_k), \\ \hat{\tilde{y}} &= C_0[\hat{x}, \hat{\eta}] + CF^a(t_k)\hat{\theta}, \\ \dot{\hat{\theta}} &= \mathcal{P}_{\Theta^a}\{\gamma(C_0\Omega + CF^a(t_k)))^T\}(\tilde{y} - \hat{\tilde{y}}), \end{cases} \tag{29}$$

where $\Omega \triangleq [\Omega_x^T, \Omega_\eta^T]^T$, the initial conditions are chosen as $\hat{x}(T_d) = 0$, $\hat{\eta}(T_d) = 0$, $\Omega_x(T_d) = 0$ and $\Omega_\eta(T_d) = 0$. The gains $L_x \in \mathbb{R}^{n_p \times n_y}$ and $L_e \in \mathbb{R}^{n_y \times n_y}$ are to be specified later. The scalar $\gamma > 0$ is the learning rate, and the projection operator $\mathcal{P}$ restricts the parameter estimate $\hat{\theta}$ to a predefined convex region Θ^a where Θ^a is given in Assumption 1.

Remark 3. *It should be noted that t_k is used in the estimator $\mathcal{E}$, which indicates that t_k^a is actually used in the attack case and t_k^f is actually used in the fault case. Such a design can eliminate the effects of the different event-triggering time instants in the attack case and in the fault case on the estimator $\mathcal{E}$. It is also worth pointing out that the distribution matrix F^a in the attack case is used to match the system structure of $\mathcal{W}_a$ such that estimation results are better in the attack case.* ∇

We now turn to investigate the stability in both of the considered threat scenarios (i.e., the replay attack scenario and the sensor fault scenario), and only investigate the learning capability in the replay attack case. We start by defining the estimation errors. By letting

$$\hat{z}_1(t) \triangleq \hat{x}(t) - \Omega_x(t)\hat{\theta}(t),\ \hat{z}_2(t) \triangleq \hat{\eta}(t) - \Omega_\eta(t)\hat{\theta}(t), \tag{30}$$

and considering both of the threat scenarios, we define the following estimation errors:

$$e_x \triangleq x - \Omega_x\theta - \hat{z}_1,\ e_\eta \triangleq \eta - \Omega_\eta\theta - \hat{z}_2,$$
$$e_y \triangleq \tilde{y} - \hat{\tilde{y}},\ \bar{e}_y \triangleq Ce_x + e_\eta,\ \tilde{\theta} \triangleq \theta - \hat{\theta}.$$

Then, from (28) and (29), the error system is obtained as

$$\dot{e}_x(t) = (A + L_xC)e_x(t) + L_xe_\eta(t), \tag{31a}$$

$$\dot{e}_\eta(t) = (-CA + L_eC)e_x(t) + L_ee_\eta(t), \tag{31b}$$

$$e_x^+(t_k) = e_x(t_k),\ e_\eta^+(t_k) = 0, \tag{31c}$$

$$e_y(t) = \bar{e}_y(t) + C_0\Omega(t)\tilde{\theta}(t) + C(F(t_k)\theta - F^a(t_k)\hat{\theta}(t)), \tag{31d}$$

$$\dot{\tilde{\theta}}(t) = -\mathcal{P}_{\Theta^a}\{\gamma(C_0\Omega(t) + CF^a(t_k)))^T\}e_y(t). \tag{31e}$$

The error system (31a)-(31c) is a time-dependent impulsive dynamical system. The stability theory of impulsive systems in Haddad et al. [12] will be exploited to investigate the stability and learning properties of the estimator $\mathcal{E}$ in (29).

Theorem 1. *Consider the system (7) potentially subject to the threat cases satisfying Assumption 3 with the replay attack and the sensor faults satisfying Assumptions 1 and 2 respectively, and consider also the estimator $\mathcal{E}$ in (29).*
(i) *If the gains L_x and L_e are designed such that there exists a matrix $P = P^T > 0$ satisfying*

$$A_0^TP + PA_0 + \alpha P < 0, \tag{32}$$

$$J^TPJ - P \leq 0, \tag{33}$$

where $\alpha > 0$ is any scalar and

$$A_0 = \begin{bmatrix} A + L_xC & L_x \\ -CA + L_eC & L_e \end{bmatrix},\ J = \begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix},$$

then the estimator $\mathcal{E}$ in (29) can guarantee that in both of the considered threat scenarios, the estimation errors $x - \hat{x}$, $\eta - \hat{\eta}$, e_y and $\tilde{\theta}$ are uniformly bounded.
(ii) *There exists a bounded function $\bar{\kappa}_1(t) > 0$ such that*

$$|\bar{e}_y(t)| \leq \bar{\kappa}_1(t),\ \forall\, t \geq T_d, \tag{34}$$

where $\bar{\kappa}_1(t)$ is given by

$$\bar{\kappa}_1(t) \triangleq |C_0|\sqrt{w(t)/\lambda_{\min}(P)}. \tag{35}$$

In the above equations, $w(t)$ is given by

$$w(t) = k_0e^{-\alpha(t-T_d)},\ \forall\, t \geq T_d, \tag{36}$$

where $k_0 \geq \lambda_{\max}(P)(|x(T_d)|^2 + |\eta(T_d)|^2)$.
(iii) *In the replay attack case, there exist a constant $\kappa_2 > 0$ and a bounded function $\zeta(t)$ such that for any finite time $t_m \geq T_d$, the output estimation error e_y satisfies*

$$\int_{T_d}^{t_m} |e_y(\tau)|^2\, dt \leq \kappa_2 + 2\int_{T_d}^{t_m} |\zeta(\tau)|^2\, dt, \tag{37}$$

where $\zeta(t)$ is specified later. ■

Proof. **(i)** Let $\xi(t) \triangleq [e_x^T(t), e_\eta^T(t)]^T$. Then, the dynamics of $e_x(t)$ and $e_\eta(t)$ in (31a), (31b) and (31c) can be written in the following compact form

$$\dot{\xi}(t) = A_0\xi(t),\ \forall\, t \in \Gamma_k, \tag{38a}$$

$$\xi^+(t) = J\xi(t),\ \forall\, t = t_{k+1}, \tag{38b}$$

where $\xi(T_d) = [e_x^T(T_d), e_\eta^T(T_d)]^T = [x^T(T_d), \eta^T(T_d)]^T$. Let $V_1(\xi(t)) = \xi^T(t)P\xi(t)$ be a Lyapunov function candidate. Then, by using the condition (32), the time derivative of V_1 in the time interval Γ_k along the solution of (38a) satisfies

$$\begin{aligned}\dot{V}_1 =& \xi^T(t)(A_0^T P + PA_0)\xi(t)\\ \le& -\alpha\lambda_{\min}(P)|\xi(t)|^2 < 0,\ \forall\ |\xi(t)| \neq 0,\ t \in \Gamma_k.\end{aligned} \tag{39}$$

Based on the condition (33) and the jump given in (38b), we have

$$V_1(\xi^+(t)) - V_1(\xi(t)) \le 0,\ \forall\ t = t_{k+1}. \tag{40}$$

Thus, according to Theorem 2.8 in [12], $\xi(t) \in L_\infty$ and thus, $e_x(t),\ e_\eta(t) \in L_\infty$. Due to the parameter projection, $\hat{\theta}(t) \in L_\infty$. In addition, (29) indicates that $\Omega_x(t), \Omega_\eta(t) \in L_\infty$. From (30), based on the definition of $e_x(t)$ and $e_\eta(t)$, we can conclude that $x(t) - \hat{x}(t) \in L_\infty$, $\eta(t) - \hat{\eta}(t) \in L_\infty$ and $\bar{e}_y(t) \in L_\infty$. Furthermore, according to Assumptions 1 and 2, $F(t_k) \in L_\infty$ and $\theta \in L_\infty$. Hence, it follows from (31d) that $e_y \in L_\infty$.

(ii) It follows from (39) and (40) that

$$\begin{aligned}\dot{V}_1(\xi(t)) =& \xi^T(t)(A_0^T P + PA_0)\xi(t)\\ \le& -\alpha\xi^T(t)P\xi(t) = -\alpha V_1(\xi(t)),\ \forall\ t \ge T_d.\end{aligned}$$

Based on the comparison principle given in [15], we can obtain that $V_1(\xi(t)) \le w(t)$ for any $t \ge T_d$ where

$$\dot{w}(t) = -\alpha w(t),\ w(T_d) \ge \lambda_{\max}(P)|\xi(T_d)|^2.$$

Recalling $\xi(T_d) = [x^T(T_d), \eta^T(T_d)]^T$, $w(t)$ and k_0 in (36) can be obtained. Furthermore, it follows from the fact that $|\xi(t)| \le \sqrt{V_1(t)/\lambda_{\min}(P)}$ and $\bar{e}_y = Ce_x + e_\eta = C_0\xi$, we can obtain

$$|\bar{e}_y(t)| \le |C_0\xi(t)| \le |C_0|\sqrt{w(t)/\lambda_{\min}(P)}.$$

Thus, $\bar{\kappa}_1(t)$ in (35) can be obtained, and result (ii) follows.

(iii) By splitting $\xi(t)$ as $\xi(t) = \xi_1(t) + \xi_2(t)$ for $t > T_d$, it follows from (38a) and (38b) that

$$\dot{\xi}_j(t) = A_0\xi_j(t),\ \forall\ t \in \Gamma_k, \tag{41a}$$

$$\xi_j^+(t) = J\xi_j(t),\ \forall\ t = t_{k+1},\ j = 1, 2, \tag{41b}$$

where $\xi_1(T_d) = 0$ and $\xi_2(T_d) = \xi(T_d)$. Also, from (31d), e_y in the attack case satisfies

$$e_y(t) = C_0(\xi_1(t) + \xi_2(t)) + (C_0\Omega(t) + CF^a(t_k^a))\tilde{\theta}(t).$$

Considering the following Lyapunov function candidate

$$V_2(\tilde{\theta}(t), \xi_2(t)) = \frac{1}{2\gamma}\tilde{\theta}^T(t)\tilde{\theta}(t) + \int_t^\infty |C_0\xi_2(\tau)|^2 d\tau,$$

the time derivative of V_2 along the solution of (31e) and (41a) is given by

$$\begin{aligned}\dot{V}_2 =& -\frac{1}{\gamma}\tilde{\theta}^T(t)\dot{\hat{\theta}}(t) + |C_0\xi_2(t)|^2\\ =& \frac{1}{\gamma}\tilde{\theta}^T(t)\mathcal{P}_{\Theta^a}\{\gamma(C_0\Omega(t) + CF^a(t_k^a))\}^T e_y(t) + |C_0\xi_2(t)|^2.\end{aligned}$$

By using the definition of the projection operator $\mathcal{P}_{\Theta^a}$ and following the logic in [8], we have

$$\begin{aligned}&\frac{1}{\gamma}\tilde{\theta}^T\mathcal{P}_{\Theta^a}\{\gamma(C_0\Omega + CF^a(t_k^a))\}^T e_y\\ &\qquad\le \tilde{\theta}^T\{(C_0\Omega + CF^a(t_k^a))\}^T e_y.\end{aligned}$$

Hence, by completing the squares, it yields

$$\dot{V}_2 \le -\frac{|e_y(t)|^2}{2} + |C_0\xi_1(t)|^2,\ \forall\ t \in \Gamma_k^a.$$

Moreover, at the jump time instant, it follows from (31c) and (41b) that

$$\begin{aligned}&V_2(\tilde{\theta}^+(t), \xi_2^+(t)) - V_2(\tilde{\theta}(t), \xi_2(t))\\ &\quad= \int_t^\infty |C_0J\xi_2^+(\tau)|^2 - |C_0\xi_2(\tau)|^2 d\tau\\ &\quad= -\int_t^\infty |C_0(I - J)\xi_2(\tau)|^2 d\tau \le 0,\ \forall\ t = t_{k+1}^a.\end{aligned}$$

Thus, by letting $\zeta(t) = |C_0\xi_1(t)|$, we can deduce that for any $t_m \ge T_d$,

$$V_2(t_m) - V_2(T_d) \le \int_{T_d}^{t_m} \frac{|e_y(\tau)|^2}{2} dt + \int_{T_d}^{t_m} |\zeta(\tau)|^2\ dt.$$

Due to the boundedness of $\xi(t)$ and $\tilde{\theta}(t)$, we can conclude $\xi_1(t)$ and $\zeta(t)$ are also bounded. Therefore, by letting $\kappa_2 \triangleq \sup_{t_m \ge T_d}(V_2(T_d) - V_2(t_m))$, the inequality (37) follows. □

The existence of k_0 satisfying $k_0 \ge \lambda_{\max}(P)(|x(T_d)|^2 + |\eta(T_d)|^2)$ in (36) can be guaranteed by Assumptions 1 and 2. In the presence of any type of threats, the replay attack satisfying Assumption 1 or the sensor bias fault satisfying Assumption 2, $x(T_d)$ and $\eta(T_d)$ can be guaranteed to be bounded, thereby guaranteeing the existence of k_0. In addition, the value of k_0 has few effects on the final threat discrimination since $w(t)$ in (36) converges to zero asymptotically.

3.3. Backward-in-time Signal Processing Model

In this section, a backward-in-time signal processor is proposed and analyzed separately in the attack scenario and in the fault scenario. We start by reconstructing $\Delta x_i(t)$, $i \in \{a, f\}$ based on the estimates provided by the estimator $\mathcal{E}$. According to $\hat{z}_1$ in (29) and $\hat{z}_2$ in (30) provided by the estimator $\mathcal{E}$, $C\hat{z}_1 + \hat{z}_2$ can be considered as a reconstruction of $\tilde{y}'$ in the attack case since $\tilde{y}_a = \tilde{y}'$, and $\hat{z}_1$ and $\hat{z}_2$ are the estimates of x' and η' respectively. In the fault case, $C\hat{z}_1 + \hat{z}_2$ can be considered as a reconstruction of $\tilde{y}$ in the nominal case since $\hat{z}_1$ and $\hat{z}_2$ are the estimates of x_n and η_n in the nominal case respectively. Hence, based on the definitions of $\Delta\tilde{y}_a$ in (20) and $\Delta\tilde{y}_f$ in (22), e_x and e_η defined after (30), a unified reconstruction for both of $\Delta\tilde{y}_a$ and $\Delta\tilde{y}_f$, denoted by $\Delta\hat{\tilde{y}}$, is proposed as

$$\Delta\hat{\tilde{y}}(t) \triangleq \tilde{y}(t) - (C\hat{z}_1(t) + \hat{z}_2(t))$$

$$\begin{aligned} &= (C_0\Omega(t) + CF(t_k))\theta + (Ce_x(t) + e_\eta(t)) \\ &= (C_0\Omega(t) + CF(t_k))\theta + \bar{e}_y(t), \end{aligned} \tag{42}$$

where $\Delta\hat{\tilde{y}}$ is a reconstruction of $\Delta\tilde{y}_a$ in the replay attack case and is a reconstruction of $\Delta\tilde{y}_f$ in the sensor bias fault case. Moreover, a unified reconstruction for both of Δx_a and Δx_f is proposed based on the estimator $\mathcal{E}$. It follows from (24) and (25) in Proposition 1 and $\theta^a = \Delta x_a(T_d)$ that $\Delta x_a(t) = e^{A(t-T_d)}\theta^a + \int_{T_d}^{t} e^{A(t-\tau)}BK\Delta\tilde{y}_a(\tau)d\tau$. Thus, by using $\hat{\theta}$ and $\Delta\hat{\tilde{y}}$, the unified reconstruction for $\Delta x_a(t)$ and $\Delta x_f(t)$, denoted by $\Delta\hat{x}(t)$, is proposed as

$$\Delta\hat{x}(t) = e^{A(t-T_d)}\hat{\theta}(t) + \int_{T_d}^{t} e^{A(t-\tau)}BK\Delta\hat{\tilde{y}}(\tau)d\tau. \tag{43}$$

It should be mentioned that $\Delta\hat{x}$ may reconstruct Δx_a well whereas it can not be guaranteed to be able to reconstruct Δx_f well enough.

The backward-in-time signal processor is proposed in the sequel. To this end, we introduce the concept of backward-in-time models. Given a forward-in-time model driven by a time-varying signal, the corresponding backward-in-time model is driven backwards in time, starting from a "terminal" state, by a particularly designed time-varying signal, such that the states of the backward-in-time model equal those of the forward-in-time model. Backward-in-time models for Markovian processes have been defined in several papers such as [17, 26], which are extended in the sequel to obtain the backward-in-time model for the deterministic process $\Delta\mathcal{W}_a$.

To describe the time running backwards, a time variable t_b running reversely from t is defined and indicated as $t_b|t$. Then, a variable x running backwards in time and starting from t can be denoted by $x(t_b|t)$. Given $\Delta\mathcal{W}_a$ in (21), the backward-in-time signal processor is proposed as the backward-in-time model of $\Delta\mathcal{W}_a$, i.e.,

$$\mathcal{B}: \left\{ \begin{array}{rcl} \dfrac{dx_b(t_b|t)}{dt_b} & = & Ax_b(t_b|t), \\ \rho(t_b|t) & = & Cx_b(t_b|t),\ \forall\, t \geq T_d, \end{array} \right. \tag{44}$$

where $x_b(t_b|t) \in \mathbb{R}^{n_p}$ is the state at the time t_b, $\rho(t_b|t) \in \mathbb{R}^{n_y}$ is the output and acts as the discrimination residual. The model $\mathcal{B}$ is activated at time instant T_d, and proceeds backwards to the time t_b. The initial condition (i.c.) of the state vector is $x_b(t|t)$ and is chosen as the reconstruction of $\Delta x(t)$, i.e.,

$$x_b(t|t) \triangleq \Delta\hat{x}(t), \tag{45}$$

where $\Delta\hat{x}(t)$ is given in (43).

The output $\rho(t_b|t)$ under the i.c. (45) is explicitly given in the following. From (43) and by solving the differential equation in (44), we have

$$\rho(t_b|t) = Ce^{A(t_b-T_d)}\hat{\theta}(t) + Cg(t, \Delta\hat{\tilde{y}}), \tag{46}$$

where

$$g(t, \Delta\hat{\tilde{y}}) \triangleq \int_{T_d}^{t} e^{A(t_b-\tau)}BK\Delta\hat{\tilde{y}}(\tau)d\tau. \tag{47}$$

Then, the threat discrimination approach based on the boundedness properties of $\rho(t_b|t)$ is shown in the following theorem.

Theorem 2. *Consider the system (7) subject to the threats satisfying Assumption 3 with the replay attack and the sensor faults satisfying Assumptions 1 and 2 respectively, and consider also the estimator $\mathcal{E}$ in (29). Then, we have the following results:*
(i) *In the replay attack case, the output $\rho(t_b|t)$ of the backward-in-time signal processor $\mathcal{B}$ in (44) with the i.c. (45) is bounded as follows:*

$$|\rho(t_b|t)| \leq J_{th}(t_b, T_d, t, t_k^a),\ \forall\, t \geq T_d, \tag{48}$$

where J_{th} is given by

$$J_{th} \triangleq k_1\sigma_a e^{-\lambda(t_b-T_d)} + k_1k_2 \int_{T_d}^{t} e^{-\lambda(t_b-\tau)}\kappa_a(\tau, t_k^a)d\tau. \tag{49}$$

In the above equation, $k_1 > 0$ and $\lambda > 0$ satisfy $|Ce^{A(t_b-T_d)}| \leq k_1 e^{-\lambda(t_b-T_d)}$, $k_2 \triangleq |CBK|$ and

$$\kappa_a(t, t_k^a) \triangleq \bar{\kappa}_1(t) + |C_0\Omega(t) + CF^a(t_k^a)|\sigma_a,$$

where $\bar{\kappa}_1(t)$ is given in (35).
(ii) *The occurrence of a replay attack is excluded if for a fixed time $t_b < T_d$, there exists a time instant $t_f > T_d$ such that $|\rho(t_b|t_f)| > J_{th}(t_b, T_d, t_f, t_k^f)$. Then, the occurring threat type is guaranteed to be the sensor bias fault(s), and the sufficient discrimination condition is*

$$\begin{aligned} &|Ce^{A(t_b-T_d)}\theta_f + Cg(t_f, \Delta\tilde{y}_f)| \\ &\quad \geq J_{th}(t_b, T_d, t_f, t_k^f) + k_1(\sigma_a + \sqrt{n_p})e^{-\lambda(t_b-T_d)} \\ &\quad + k_1k_2 \int_{T_d}^{t_f} e^{-\lambda(t_b-\tau)}\bar{\kappa}_1(\tau)d\tau, \end{aligned} \tag{50}$$

where $\bar{\kappa}_1(t)$ is given in (35). ■

Proof. **(i)** According to [28], for the Hurwitz matrix A, there exist $\lambda > 0$ and $k_1 > 0$ such that $|Ce^{A(t_b-T_d)}| \leq k_1 e^{-\lambda(t_b-T_d)}$. Thus, from $|\hat{\theta}| \leq \sigma_a$ due to the projection operator $\mathcal{P}_{\Theta^a}$, we can obtain

$$|e^{A(t_b-T_d)}\hat{\theta}(t)| \leq k_1\sigma_a e^{-\lambda(t_b-T_d)}.$$

Moreover, based on (34) in Theorem 1 and Assumption 1, and using the triangle inequality, $\Delta\hat{\tilde{y}}$ in (42) in the attack case satisfies

$$|\Delta\hat{\tilde{y}}(t)| \leq \kappa_a(t, t_k^a).$$

Thus, based on the definition of g in (47) and ρ in (46), we can derive

$$|\rho(t_b|t)| \leq k_1\sigma_a e^{-\lambda(t_b-T_d)} + k_1k_2 \int_{T_d}^{t} e^{-\lambda(t_b-\tau)}\kappa_a(\tau, t_k^a)d\tau.$$

Hence, J_{th} in (49) can be obtained and result **(i)** follows.
(ii) In the fault case, $\Delta\tilde{y}_f(t) \neq 0$. From (43) and the initial condition (45), we have

$$\begin{aligned}\rho(t_b|t) = &\, Ce^{A(t_b-T_d)}(\theta^f - \tilde{\theta}(t)) \\ &+ Cg(t, \Delta\tilde{y}_f) - Cg(t, \Delta\tilde{y}_f - \Delta\hat{\tilde{y}}).\end{aligned}$$

Using the reverse triangle inequality, we have

$$\begin{aligned}|\rho(t_b|t)| \geq &\, |Ce^{A(t_b-T_d)}\theta^f + Cg(t, \Delta\tilde{y}_f)| \\ &- |Ce^{A(t_b-T_d)}\tilde{\theta}(t)| - |Cg(t, \Delta\tilde{y}_f - \Delta\hat{\tilde{y}})|.\end{aligned}$$

By considering $Cz_1 + z_2$ as $\tilde{y}_n$, it follows from (42) that in the fault case, $\Delta\tilde{y}_f - \Delta\hat{\tilde{y}} = \bar{e}_y(t)$. Thus, it follows from (34) in Theorem 1 that in the fault case, $|\Delta\tilde{y}_f - \Delta\hat{\tilde{y}}| \leq \bar{\kappa}_1(t)$. From the definition of g in (47), we have

$$|Cg(t, \Delta\tilde{y}_f - \Delta\hat{\tilde{y}})| \leq k_1 k_2 \int_{T_d}^{t} e^{-\lambda(t_b-\tau)}\bar{\kappa}_1(\tau)d\tau.$$

In addition, it follows from $\theta^f = [1, 1, \cdots, 1]^T \in \mathbb{R}^{n_p}$ that $|\theta^f| \leq \sqrt{n_p}$. Additionally, $|\hat{\theta}| \leq \sigma_a$ due to the projection operator $\mathcal{P}_{\Theta^a}$. Thus, in the fault case $|\tilde{\theta}(t)| \leq |\hat{\theta}| + |\theta^f| \leq \sigma_a + \sqrt{n_p}$, and further, for the Hurwitz matrix A, we have

$$|Ce^{A(t_b-T_d)}\tilde{\theta}(t)| \leq k_1(\sigma_a + \sqrt{n_p})e^{-\lambda(t_b-T_d)}.$$

In order to exclude the replay attack, the inequality $|\rho(t_b|t)| > J_{th}(t_b, T_d, t, t_k^f)$ must hold at some time $t_f > T_d$. Therefore, the sufficient condition (50) can be obtained, and a sensor bias fault is identified if (50) holds for some time $t_f > T_d$.□

Based on Theorem 2, the fault discrimination principle is given as follows: if there exists a time t_f such that $|\rho(t_b|t_f)| > J_{th}(t_b, T_d, t_f, t_k^f)$, then the detected threat is identified as sensor bias fault(s) (since the attack case is excluded). However, in the case that $|\rho(t_b|t)| \leq J_{th}(t_b, T_d, t, t_k^a)$ or $|\rho(t_b|t)| \leq J_{th}(t_b, T_d, t, t_k^f)$ for any $t \geq T_d$, the threat may be either a replay attack or sensor fault(s), and thus, no decision regarding the type of threat (attack or sensor fault) can be made.

Remark 4. *The limitations of the developed methodology in this paper are discussed in the sequel. One limitation of the developed methodology is that it can not distinguish the replay attack from the considered two threat scenarios (i.e., the replay attack scenario and the sensor bias fault scenario). In order to be able to identify the replay attacks as well, an additional signal processing model is required to be designed based on the characteristics of the incremental system $\Delta\mathcal{W}_f$ (due to the sensor bias fault(s)). Another limitation is the design conditions for the observer gains L_x and L_e given in result (i) of Theorem 1. Theses conditions, specifically (32) and (33), are somewhat conservative in terms of guaranteeing the stability of the impulsive error system (31a)-(31c), which results in the fact that the required L_x and L_e are hard to be found or even do not exist for some practical systems. Hence, additional work is required to relax conditions (32) and (33) for the stability of the impulsive system (31a)-(31c).* ∇

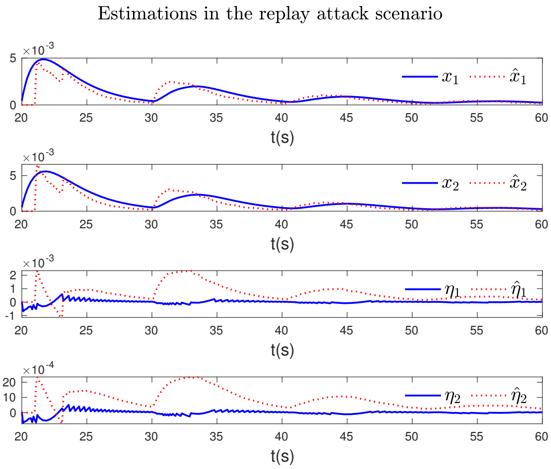

Figure 2: The estimates of the states x and the estimates of the transmission errors η in the attack scenario.

4. Simulation

In this section, a numerical simulation is presented. The matrices of the CPS described in (7) are given as follows

$$A = \begin{bmatrix} -1 & 0.1 \\ 0.3 & -0.99 \end{bmatrix}, \ B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \ C = I_2.$$

The feedback gain is given by $K = [0.1286, 0.2614]$. In addition, the parameter δ in the event-triggering condition given in (3) is set as $\delta = 0.2$. Regarding the potential threats with the replay attack satisfying Assumption 1 and the sensor bias faults satisfying Assumption 2, σ_a in Assumption 1 and σ_f in Assumption 2 are given by $\sigma_a = 3$ and $\sigma_f = 40$ respectively.

The adaptive estimator is constructed based on $\mathcal{E}$ given in (29) where the learning rate γ is set as $\gamma = 400000$, and the gain matrices L_x and L_e are calculated based on Theorem 1 as

$$L_x = \begin{bmatrix} -3.1661 & 0.1563 \\ -0.0233 & -3.6866 \end{bmatrix} \times 10^4, \ L_e = \begin{bmatrix} -1.8789 & -0.0020 \\ 0.2546 & -1.5917 \end{bmatrix} \times 10^4.$$

In order to construct the threshold J_{th} given in (49) in Theorem 2, we choose $k_1 = 1.5$, $k_2 = 0.5$ and $\lambda = -0.82$. In addition, we assume that a threat is detected at $T_d = 21$s, and the time t_b is set as $t_b = 20$s. Thus, the residual is constructed based on (46), and the threshold is constructed based on (49).

For the simulation purpose, the details of the replay attack event is given. In the attack scenario, the attacker records the sensor measurements from 10s to 20s, and then replays it from 20s to 30s. The estimation results of the estimator $\mathcal{E}$ and the event-triggering time instants t_k^a in the replay attack case are shown in Fig. 2 and 3. As it can be seen from Fig. 2, the estimates of the states x and the estimates of the transmission errors η converge to the real values asymptotically. The attack reconstruction $\hat{a}(t)$ in Fig. 3 also converges to $a(t)$ asymptotically. The residual in (46) and the threshold in (49) are constructed using the transmission time instants t_k^a given in Fig. 3. The threat discrimination results in the

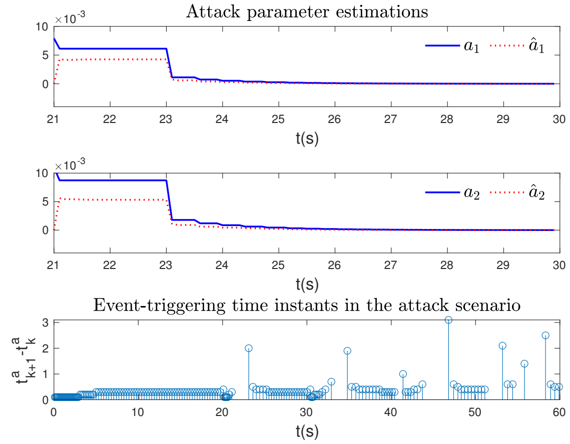

Figure 3: The reconstruction $\hat{a}(t) = F_a(t_k^a)\hat{\theta}(t)$ of the virtual attack signal $a(t)$ and the event-triggering time instants t_k^a in the attack scenario.

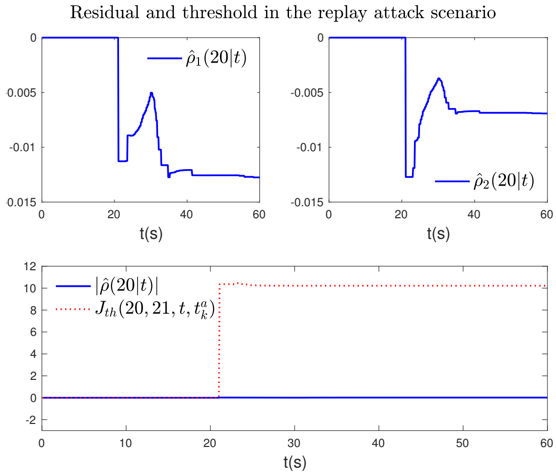

Figure 4: The residuals and the threshold in the attack scenario.

replay attack scenario are shown in Fig. 4. The third subfigure illustrates that for $t > T_d$, $|\rho(t_b|t)|$ (blue solid line) remains below the threshold $J_{th}(t_b, T_d, t, t_k^a)$ (red dot line), which are in line with the theoretical finding (i) in Theorem 2. However, a decision regarding the type of the threat (i.e., an attack is occurring) can not be made.

In the sensor fault scenario, the sensor bias fault signal $f(t) = [-20, 20 - 12\sin(t)]^T$ is considered to occur and present for $t \geq 20$s. The estimations provided by the estimator $\mathcal{E}$ and the event-triggering time instants t_k^f in the presence of the sensor bias faults are shown in Figs. 5 and 6. As it can be seen from Fig. 5, the estimates of the states x and the estimate of the transmission errors η are bounded in the presence of the sensor faults, and also, the fault reconstruction $\hat{f}(t)$ in Fig. 5 is bounded, but can not converge to $f(t)$. Note that the estimates provided by the estimator $\mathcal{E}$ in the fault case are not good, which satisfies our expectation since $\mathcal{E}$ is particularly designed for the attack case. This creates sufficiently discrepancy to allow the identification of the sensor fault case. The residual in (46) and the threshold in (49) are constructed using the transmission time instants t_k^f given in Fig. 6. The threat discrimination results in the sensor bias fault scenario are shown in Fig. 7, in which the third sub-figure illustrates that at about $t_f = 21$s, $|\rho(t_b|t_f)|$ (blue solid line) exceeds the threshold $J_{th}(t_b, T_d, t_f, t_k^f)$ (red dot line). Hence, at around $t_f = 21$s, the attack case is excluded and the threat is identified to be the case of a type of sensor bias fault(s).

5. Conclusions

Using an event-triggered data communication, a threat discrimination methodology has been proposed for CPS to identify the sensor bias fault(s) case between two threat cases: replay attack and sensor bias fault(s). Distinct incremental systems due to the replay attack and the sensor fault(s) have been established. An event-triggered adaptive estimator has been designed to reconstruct the unknown increments, and a backward-in-time model has been proposed for utilizing the difference between the incremental systems to generate distinguishable outputs. The threat discrimination condition was rigorously investigated and characterizes the class of distinguishable sensor bias faults. In the future, we will focus on developing a comprehensive scheme for identifying both threat scenarios, and also on developing a unified framework, allowing to discriminate between general cyber attacks and physical faults, rather than between the specific replay attacks and sensor bias faults.

References

[1] Barboni, A., Rezaee, H., Boem, F., Parisini, T., 2020. Detection of covert cyber-attacks in interconnected systems: a distributed model-based approach. IEEE Transactions on Automatic Control 65, 3728–3741.

[2] Blanke, M., Kinnaert, M., Lunze, J., Staroswiecki, M., Schröder, J., 2006. Diagnosis and fault-tolerant control. Springer Science & Business Media.

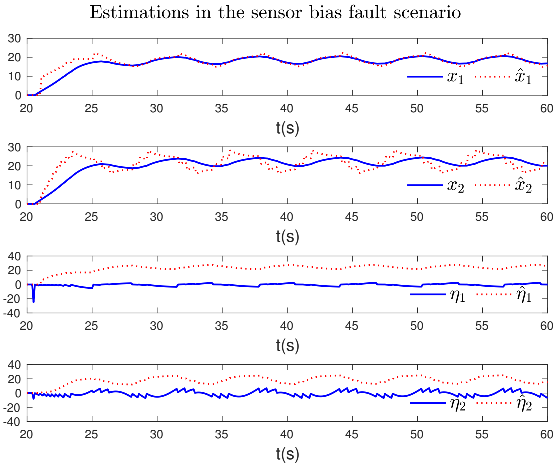

Figure 5: The estimates of the states x and the estimation of the transmission errors η in the fault scenario.

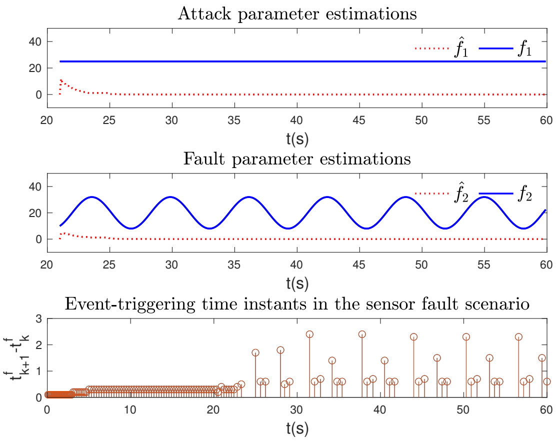

Figure 6: The reconstruction $\hat{f}(t) = F_a(t_k^f)\hat{\theta}(t)$ of the fault parameter signal $f(t)$ and the event-triggering time instants t_k^f in the fault scenario.

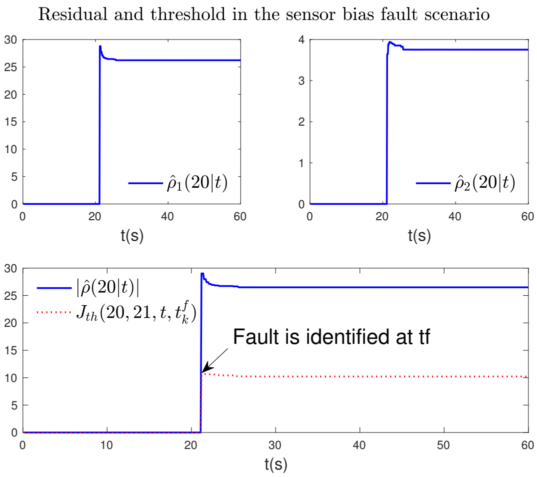

Figure 7: The residuals and the threshold in the fault scenario.

[3] Cardenas, A., Amin, S., Sastry, S., 2008. Secure control: Towards survivable cyber-physical systems, in: the 28th International Conference on Distributed Computing Systems Workshops, IEEE. pp. 495–500.

[4] Chanthery, E., Subias, A., 2019. Diagnosis approaches for detection and isolation of cyber attacks and faults on a two-tank system, in: the 30th International Workshop on Principles of Diagnosis DX'19.

[5] Dibaji, S., Pirani, M., Flamholz, D., Annaswamy, A., Johansson, K., Chakrabortty, A., 2019. A systems and control perspective of cps security. Annual Reviews in Control 47, 394–411.

[6] Ding, D., Han, Q., Xiang, Y., Ge, X., Zhang, X., 2018. A survey on security control and attack detection for industrial cyber-physical systems. Neurocomputing 275, 1674–1683.

[7] Ding, S., 2008. Model-based fault diagnosis techniques: design schemes, algorithms, and tools. Springer Science & Business Media.

[8] Farrell, J., Polycarpou, M., 2006. Adaptive approximation based control: unifying neural, fuzzy and traditional adaptive approximation approaches. John Wiley & Sons.

[9] Ferrari, R., Teixeira, A., 2017. Detection and isolation of replay attacks through sensor watermarking. IFAC-Papers OnLine 50, 7363–7368.

[10] Ferrari, R., Teixeira, A., 2020. A switching multiplicative watermarking scheme for detection of stealthy cyber-attacks. IEEE Transactions on Automatic Control .

[11] Fidler, D.P., 2011. Was stuxnet an act of war? decoding a cyberattack. IEEE Security & Privacy 9, 56–59.

[12] Haddad, W., Chellaboina, V., Nersesov, S., 2006. Impulsive and hybrid dynamical systems: stability, dissipativity, and control. Princeton University Press.

[13] Heemels, W., Johansson, K., Tabuada, P., 2012. An introduction to event-triggered and self-triggered control, in: the 51st IEEE Conference on Decision and Control (CDC), pp. 3270–3285.

[14] Keijzer, T., Ferrari, R., 2019. A sliding mode observer approach for attack detection and estimation in autonomous vehicle platoons using event triggered communication, in: the 58th Conference on Decision and Control (CDC), pp. 5742–5747.

[15] Khalil, H., 2002. Nonlinear systems. Prentice hall Upper Saddle River, NJ.

[16] Liu, X., Su, X., Shi, P., Nguang, S., Shen, C., 2019. Fault detection filtering for nonlinear switched systems via event-triggered communication approach. Automatica 101, 365–376.

[17] Ljung, L., Kailath, T., 1976. Backwards Markovian models for second-order stochastic processes. IEEE Transactions on Information Theory 22, 488–491.

[18] Lunze, J., Lehmann, D., 2010. A state-feedback approach to event-based control. Automatica 46, 211–215.

[19] Mo, Y., Sinopoli, B., 2009. Secure control against replay attacks, in: the 47th annual Allerton conference on communication, control, and computing, pp. 911–918.

[20] Romagnoli, R., Weerakkody, S., Sinopoli, B., 2019. A model inversion based watermark for replay attack detection with output tracking, in: American Control Conference, pp. 384–390.

[21] Smith, R., 2011. A decoupled feedback structure for covertly appropriating networked control systems. IFAC-Paper onLine 44, 90–95.

[22] Smith, R., 2015. Covert misappropriation of networked control systems: Presenting a feedback structure. IEEE Control Systems Magazine 35, 82–92.

[23] Tabuada, P., Member, S., 2007. Event-triggered real-time scheduling of stabilizing control tasks. IEEE Transactions on Automatic Control 52, 1680–1685.

[24] Taheri, M., Khorasani, K., Shames, I., Meskin, N., 2020. Cyber attack and machine induced fault detection and isolation methodologies for cyber-physical systems , 1–10URL: http://arxiv.org/abs/2009.06196.

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

[26] Verghese, G., Kailath, T., 1979. A further note on backwards Markovian models. IEEE Transactions on Information Theory 25, 121–124.

[27] Zhang, K., Polycarpou, M., Parisini, T., 2020. Enhanced anomaly detector for nonlinear cyber-physical systems against stealthy integrity attacks, in: IFAC World Congress, p. Accepted.

[28] Zhang, X., Polycarpou, M., Parisini, T., 2010. Fault diagnosis of a class of nonlinear uncertain systems with lipschitz nonlinearities using adaptive estimation. Automatica 46, 290–299.

[29] Zhong, M., Ding, S., Zhou, D., He, X., 2020. An $H\_/H_\infty$ optimization approach to event-triggered fault detection for linear discrete-time systems. IEEE Transactions on Automatic Control 65, 4464–4471.