Passive Attack Detection for a Class of Stealthy Intermittent Integrity Attacks

Kangkang Zhang, Christodoulos Keliris, Thomas Parisini, *Fellow, IEEE*, Bin Jiang, *Fellow, IEEE*, and Marios M. Polycarpou, *Fellow, IEEE*

***Abstract*—This paper proposes a passive methodology for detecting a class of stealthy intermittent integrity attacks in cyber-physical systems subject to process disturbances and measurement noise. A stealthy intermittent integrity attack strategy is first proposed by modifying a zero-dynamics attack model. The stealthiness of the generated attacks is rigorously investigated under the condition that the adversary does not know precisely the system state values. In order to help detect such attacks, a backward-in-time detection residual is proposed based on an equivalent quantity of the system state change, due to the attack, at a time prior to the attack occurrence time. A key characteristic of this residual is that its magnitude increases every time a new attack occurs. To estimate this unknown residual, an optimal fixed-point smoother is proposed by minimizing a piece-wise linear quadratic cost function with a set of specifically designed weighting matrices. The smoother design guarantees robustness with respect to process disturbances and measurement noise, and is also able to maintain sensitivity as time progresses to intermittent integrity attack by resetting the covariance matrix based on the weighting matrices. The adaptive threshold is designed based on the estimated backward-in-time residual, and the attack detectability analysis is rigorously investigated to characterize quantitatively the class of attacks that can be detected by the proposed methodology. Finally, a simulation example is used to demonstrate the effectiveness of the developed methodology.**

***Index Terms*—Backward-in-time equivalent quantity, fixed-point smoother, intermittent integrity attacks.**

Manuscript received October 27, 2022; accepted November 11, 2022. This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skodowska-Curie (101027980 (CSP-CPS-A-ICA), 739551 (KIOS CoE-TEAMING)), the Italian Ministry for Research in the Framework of the 2017 Program for Research Projects of National Interest (PRIN) (2017YKXYXJ), the National Natural Science Foundation of China (61903188, 62073165, 62020106003), the Natural Science Foundation of Jiangsu Province (BK20190403), the 111 Project (B20007), and the Priority Academic Program Development of Jiangsu Higher Education Institutions. Recommended by Associate Editor Hongyi Li. *(Corresponding author: Kangkang Zhang.)*

Citation: K. K. Zhang, C. Keliris, T. Parisini, B. Jiang, and M. M. Polycarpou, " Passive attack detection for a class of stealthy intermittent integrity attacks," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 4, pp. 898–915, Apr. 2023.

K. K. Zhang is with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, UK, and also with the KIOS Research and Innovation Center of Excellence, University of Cyprus, Nicosia 1678, Cyprus (e-mail: kzhang5@ic.ac.uk).

C. Keliris and M. M. Polycarpou are with the KIOS Research and Innovation Center of Excellence and the Department of Electrical and Computer Engineering, University of Cyprus, Nicosia 1678, Cyprus (e-mail: keliris.chris@gmail.com; mpolycar@ucy.ac.cy).

T. Parisini is with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, UK, and with the Department of Engineering and Architecture, University of Trieste, Trieste 34127, Italy, and also with the KIOS Research and Innovation Center of Excellence, University of Cyprus, Nicosia 1678, Cyprus (e-mail: t.parisini@gmail.com).

B. Jiang is with the College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China (e-mail: binjiang@nuaa.edu.cn).

I. Introduction

CYBER-PHYSICAL systems (CPS) integrate control, computation and communication techniques with physical engineered control systems [1]. Due to the emergence of such a complex integration, more security vulnerabilities in CPS arise and more malicious cyber threats greatly endanger various key aspects of CPS operation. A series of cyber attack events, such as the Stuxnet worm attack on the Iranian nuclear facilities, the attack on the Ukrainian power distribution network, and the recent colonial oil pipeline attack in USA, have taken place in recent years (more details and examples can be found in [2]–[5]). Therefore, state-of-the-art cyber attack diagnostic technologies are required to safeguard the operation of CPS against possible malicious attacks.

A. State of the Art

Integrity in computer science refers to the trustworthiness of data, whereas in the context of CPS, integrity attacks compromise the integrity of the transmitted data of the CPS [1]. Replay attacks [6], covert attacks [7], [8], zero-dynamics attacks [9] and false-data injection attacks [10] are the most commonly studied stealthy integrity attacks. Several survey papers provide overviews of integrity attacks from a system control perspective (see, e.g., [2], [9], [11] and [12]). Recently, some new types of integrity attacks have also been investigated. For instance, in order to achieve perfect stealthiness, a self-generated approach is developed in [13] for generating particular false data, and [14] considers a class of closed-loop nonlinear systems and develops a stealthy integrity attack formulation approach for such systems. In the aforementioned integrity attack studies, false data are injected into the CPS continuously, whereas the problem of intermittent data-injection is overseen. It is important to note that the intermittent mode of attack injection may greatly affect both the stealthiness of an attack event and the power energy consumed by an attacker.

Intermittent attacks are characterized by piece-wise attack signals. Denial-of-service attacks (DOS) [15]–[18] are typical examples of intermittent attacks. The available power energy to the attacker is optimally managed by scheduling the DoS attack application time instants in [16], [17], whereas the stealthiness of the attack is not considered. In this paper, we

consider integrity attacks performed in an intermittent mode; these are referred to as intermittent integrity attacks. Unlike the strategy for typical continuous integrity attacks, the strategy for generating intermittent integrity attacks includes two steps: 1) constructing a stealthy attack model, and 2) scheduling attack-activating and attack-pausing time instants. In terms of attack models, typical stealthy integrity attack models such as zero-dynamics attack models [14], [19], replay attack models [6] and covert attack models [20], [21], can also act as models for intermittent integrity attacks since they are undetectable by typical anomaly detectors. The scheduling of intermittent attacks (activation and pausing time instants) can improve their stealthiness through activating the attacks for a limited amount of time such that anomaly detectors do not have sufficient time to detect them. Using attacks with specially scheduled activating and pausing time instants can also save the adversary's instantaneous power energy by avoiding attack signal divergence requirements. A counterexample to this is the zero-dynamics attack case, in which the divergence attack signals are used (e.g., [9]), and may consume large instantaneous power energy. Moreover, intermittent integrity attacks can overcome the attack defense strategy such as topology switching in the case of multi-agent systems [22]. Particularly, the adversary may pause the attack prior to the switch of the topology, and then resume and update the attack to maintain the stealthiness in the new topology.

In the past decade, several methods for detecting stealthy integrity attacks have been proposed by the research community, which fell into two categories: active and passive detection methods. In active detection methods, such as the watermarking and moving-target approaches, either authentication signals are injected into the information flows of CPS or secret modules are embedded to CPS loops and series-connected to physical plants (see, e.g., [23]–[27]). On the other hand, in passive attack detection methods, only analytical redundancy approaches are used to detect integrity attacks, without using any authentication signals or secret modules [28]. However, typical passive anomaly detectors are not able to provide desirable attack detection performance. Note that attack detectors are a special type of anomaly detectors specifically designed for detecting attacks. For example, fault detection schemes in [29]–[33] may not be able to detect stealthy integrity attacks (such as [6]–[9] and [13]). Generally, the reason for this is that, in the presence of a stealthy integrity attack, the information resources (sensory measurements and control inputs) of analytical redundancy approaches remain either unchanged or slightly altered and therefore, analytical redundancy approaches that are inherently not sufficiently sensitive to such slight changes, are unable to detect stealthy integrity attacks. In [34], multiple filters are combined to formulate a type of analytical redundancy-based passive attack detector, which can detect various types of stealthy integrity attacks. However, such a detector requires additional physical communication channels, which may not be feasible or realistic. Traditional analytical redundancy-based passive anomaly detectors are enhanced in [35] and [36] for detecting stealthy integrity attacks by using a backward-in-time signal processor. In these studies, even a small change due to a stealthy integrity attack is amplified by a backward-in-time signal processor such that the amplified change becomes sufficiently "large" to be detected. However, stealthy intermittent integrity attacks are not considered in [34]–[36].

The detection of stealthy intermittent integrity attacks remains an open problem and few research works have been published. For example, [37] combines an analytical redundancy-based passive detector and a set-theoretic detector for detecting intermittent integrity attacks. In this method, the transient overshoots of the analytical redundancy and the strict detection guarantees of the set-theoretic detector are integrated to detect promptly the intermittent attacks. However, the stealthiness of the attacks is not considered in [37], which prevents the detection method from applied effectively in the case of stealthy intermittent integrity attacks.

B. Main Contributions

This paper utilizes the backward-in-time approach in the context of intermittent integrity attacks, and proposes an analytical redundancy-based passive detection methodology for detecting a class of stealthy intermittent integrity attacks. Specifically, the contributions of this paper are summarized as follows:

1) A stealthy intermittent integrity attack generation strategy is formulated, which does not require that the adversary has precise knowledge of the system states. A backward-in-time detection residual is formulated, which increases in magnitude each time a new attack occurs;

2) An optimal fixed-point smoother with covariance matrix resetting is proposed to implement the aforementioned backward-in-time residual. Such a smoother guarantees robustness to both disturbances and noise, and can also reset the covariance matrix to maintain sensitivity to intermittent integrity attacks;

3) The corresponding adaptive threshold is designed, and an attack detectability analysis is carried out to characterize quantitatively the class of detectable stealthy intermittent integrity attacks.

In terms of the stealthy intermittent integrity attacks, compared to [22], the attack generation proposed in this paper addresses the practical issue that the adversary does not have precise knowledge of the system states. In addition, in contrast with [38] in which the pausing and resuming time instants of intermittent integrity attacks are scheduled for saving power energy, this paper focuses on designing the attack generation strategy such that the generated attacks are stealthy, regardless of the pausing and resuming time instants.

Compared to the authors' previous work [36], the intermittency feature of stealthy integrity attacks is considered in this paper. Moreover, the designed smoother in this paper introduces the covariance matrix resetting technique, which is shown to guarantee robustness to both disturbances and noise, and simultaneously guarantee sensitivity to stealthy intermittent integrity attacks.

C. Notations

Consider a vector signal $x(t) : \mathbb{R}_+ \to \mathbb{R}^n$. Then, $x(t) = 0$ for $t \in [t_1, t_2] \subset \mathbb{R}_+$ means that $x(t) = 0$ identically for all $t \in [t_1, t_2]$;

$x(t) \neq 0$ for $t \in [t_1, t_2] \subset \mathbb{R}_+$ means that $x(t) \neq 0$ for at least one time instant $t \in [t_1, t_2]$. 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 set S, $|S|$ represents the number of the elements in S. A vector $x(t) \in \mathcal{L}_2[t_1, t_2]$ if $\int_{t_1}^{t_2} x^T(\tau)x(\tau)d\tau$ is finite. For a signal $x(t)$ in the finite time interval $[t_1, t_2]$ and a given matrix $R \geq 0$ with proper dimensions, we define $\|x(t)\|_R^2 = \int_{t_1}^{t_2} x^T(\tau)Rx(\tau)d\tau$. For a constant vector x and a given matrix $R \geq 0$ with proper dimensions, we define $\|x\|_R^2 = x^T Rx$. For a matrix A, $\bar{\sigma}(A)$ and $\underline{\sigma}(A)$ represent the maximum and minimum singular values of A respectively. Italics in the paper are used to highlight important associated sentences and terminologies.

The rest of this paper is organized as follows. In Section II, the problem is formulated. Section III analyzes the stealthiness of the intermittent integrity attacks formulated in this work. In Section IV, the backward-in-time detection residual is introduced, and its theoretical feasibility to indicate the stealthy intermittent integrity attacks is presented. The design details of the implementable backward-in-time detection methodology are presented in Section V and the attack detectability analysis is shown in Section VI. Section VII presents a simulation example and finally, the conclusions are drawn in Section VIII.

II. Problem Formulation

A general structure of a CPS subject to integrity cyber attacks is depicted in Fig. 1. It consists of a physical plant $\mathcal{P}$, a feedback controller C, an anomaly detector $\mathcal{D}$, an actuator communication network $\mathcal{N}_a$ and a sensor communication network $\mathcal{N}_s$.

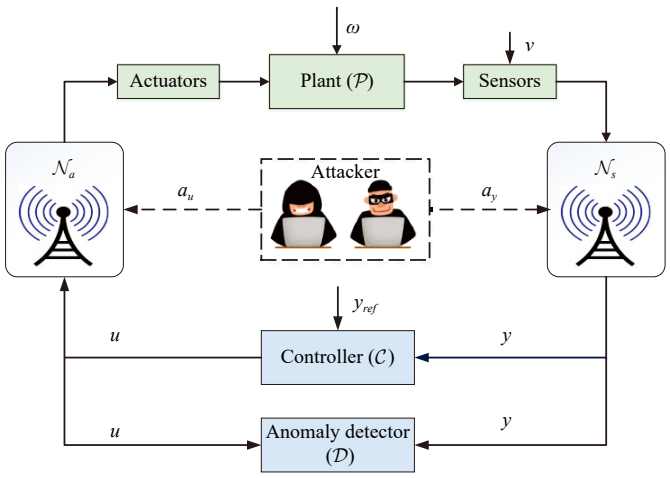


Fig. 1. Schematic diagram of CPS in the presence of integrity cyber attacks.

A. Closed-Loop CPS

In order to simplify the notation, the closed-loop CPS including C, $\mathcal{P}$, $\mathcal{N}_a$ and $\mathcal{N}_s$ are jointly denoted by $\mathcal{W}$ throughout the paper. The closed-loop system $\mathcal{W}$ is described by

$$\mathcal{W}: \begin{cases} \dot{x}(t) = Ax(t) + Bu(t) + B\Gamma_u a_u(t) + D\omega(t) \\ u(t) = Ky(t) + y_{\text{ref}}(t) \\ y(t) = Cx(t) + v(t) + \Gamma_y a_y(t) \end{cases} \tag{1}$$

where $x \in \mathbb{R}^{n_x}$ is the state vector, $u \in \mathbb{R}^{n_u}$ is the vector of control data generated by the controller C, $y \in \mathbb{R}^{n_y}$ is the vector of sensor measurements received by the controller and the anomaly detector. The integers n_x and n_y satisfy $n_y < n_x$. Moreover, $y_{\text{ref}}(t) \in \mathbb{R}^{n_y}$ denotes the output reference signals, and $\omega(t) \in \mathbb{R}^{n_\omega}$ and $v(t) \in \mathbb{R}^{n_y}$ represent the vectors of process disturbances and measurement noise, respectively. Let $K_u \subseteq \{1, \ldots, n_u\}$ and $K_y \subseteq \{1, \ldots, n_y\}$ represent the disruption resources available to the attacker, i.e., the sets of actuator and sensor communication channels respectively that can be affected by the attacker. The distribution matrices $\Gamma_u \in \mathbb{B}^{n_u \times |K_u|}$ and $\Gamma_y \in \mathbb{B}^{n_y \times |K_y|}$ ($\mathbb{B} = \{0, 1\}$) are the binary incidence matrices mapping the attack signal to the respective channels. The attack signals are $a_u(t) = [a_{u,1}(t), \ldots, a_{u,|K_u|}(t)]^T \in \mathbb{R}^{|K_u|}$ and $a_y(t) = [a_{y,1}(t), \ldots, a_{y,|K_y|}(t)]^T \in \mathbb{R}^{|K_y|}$. For each $i \in \{1, \ldots, |K_u|\}$, $a_{u,i}(t) = 0$ for $t \in \mathbb{R}_+$ if no attack occurs on the i-th transmission channel of $\mathcal{N}_a$, and similarly, for each $j \in \{1, \ldots, |K_y|\}$, $a_{y,j}(t) = 0$ for $t \in \mathbb{R}_+$ if the j-th transmission channel of $\mathcal{N}_s$ is not under attack. Throughout this paper, we denote jointly the attack signals a_u and a_y as $a(t) = [a_u^T(t), a_y^T(t)]^T$. We consider that the attack starts at an unknown time $t = T_0$ and hence $a(t) = 0$ for $0 \leq t < T_0$. By letting $B_a = [B\Gamma_u, 0_{n_x \times |K_y|}]$ and $D_a = [0_{n_y \times |K_u|}, \Gamma_y]$, then $B\Gamma_u a_u(t) = B_a a(t)$ and $\Gamma_y a_y(t) = D_a a(t)$.

In addition, the matrices $A \in \mathbb{R}^{n_x \times n_x}$, $B \in \mathbb{R}^{n_x \times n_u}$, $C \in \mathbb{R}^{n_y \times n_x}$ and $D \in \mathbb{R}^{n_x \times n_\omega}$ are system matrices known by the defender. The pair (A, D) is assumed to be stabilizable and (C, A) is assumed to be observable. The control gain $K \in \mathbb{R}^{n_u \times n_x}$ is designed to stabilize the system. In the nominal case ($a(t) = 0$ for $t \in \mathbb{R}_+$), the closed-loop system $\mathcal{W}^n$ is given by

$$\mathcal{W}^n: \begin{cases} \dot{x}^n(t) = Ax^n(t) + Bu^n(t) + D\omega(t) \\ u^n(t) = Ky^n(t) + y_{\text{ref}}(t) \\ y^n(t) = Cx^n(t) + v(t) \end{cases} \tag{2}$$

where x^n, u^n and y^n represent the state, the control input and the output, respectively, in the nominal case.

B. Anomaly Detector

We consider that the CPS is equipped with a typical anomaly detector $\mathcal{D}$ (see Fig. 1) for detecting some normal anomalies (faults or attacks). Specifically, the anomaly detector $\mathcal{D}$ contains a detection residual $r(t)$ and a constant threshold J_{th}. Without loss of generality, based on [39], we consider that the residual has the following form:

$$r(t) = \mathcal{D}(u(t), y(t), y_{\text{ref}}(t)) - y(t) \tag{3}$$

where $\mathcal{D}: \mathbb{R}^{n_u} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_y} \to \mathbb{R}^{n_y}$ is an observer (robust Luenberger observer [39], Kalman filter [29], adaptive observer [40] or sliding mode observer [41]) designed based on the analytical redundancy concept to generate an estimate of $y(t)$ in the nominal case. By using the evaluation function $J(t) = |r(t)|$, the occurrence of an anomaly (fault or attack) is ascertained if at some time $t > T_0$, $J(t)$ exceeds the threshold J_{th}, i.e.,

$$J(t) > J_{th}, \text{ alarm triggering.} \tag{4}$$

It should be noted that the residual $r(t)$ contributing to indicate the occurrence of the anomaly is based on the estimates of the system outputs at the time posterior to the anomaly

occurrence time. Hence, $\mathcal{D}$ is referred to as a forward-in-time detector in this paper. Most fault detectors in the literature such as [29], [30], [39], belong to the class of forward-in-time detectors.

C. Stealthy Intermittent Integrity Attacks

Throughout this work, to simplify the presentation, stealthiness is used to refer to stealthiness with respect to the typical standard anomaly detector $\mathcal{D}$ characterized by the residual r in (3) and the threshold J_{th} in (4). The attacker has the following available resources: partial model knowledge (A, $B\Gamma_u$, C) and disruption resources (K_y, K_u), where $B\Gamma_u$ represents the partial columns of B corresponding to K_u. The strategy for generating stealthy intermittent integrity attacks includes two steps: 1) constructing attack models, and 2) scheduling attack-activating and attack-pausing time instants. For Step 1), the model used for generating stealthy integrity attack is given in Section III. For Step 2), in this paper, we consider that the activating and pausing time instants have been scheduled by the attacker, and the attack model is activated at the following time instants:

$$t_1, \ldots, t_{N_a},\ N_a \in \mathbb{N}_+.$$

In addition, we consider that the k-th attack is active for a time length τ_k (dwell time) where $0 < \tau_k \le t_{k+1} - t_k$ and hence, it is inactive (attack silence) for $t \ge t_k + \tau_k$. Then, the activating time interval Ω_k^{ac} (attack active) and the silence time interval Ω_k^{si} (attack silence) for the k-th attack can be given respectively as follows:

$$\Omega_k^{\text{ac}} = [t_k, t_k + \tau_k),\ \Omega_k^{\text{si}} = [t_k + \tau_k, +\infty). \tag{5}$$

Also, we define an action time interval Ω_k^0 that the k-th attack affects the system as follows:

$$\Omega_k^0 = \Omega_k^{\text{ac}} \cup \Omega_k^{\text{si}} = [t_k, +\infty). \tag{6}$$

In addition, an auxiliary time interval "attack slot Ω_k" is defined as follows:

$$\Omega_k = [t_k, t_{k+1}),\ \forall\, k \in \{1, \ldots, N_a - 1\},\ \Omega_{N_a} = [t_{N_a}, +\infty). \tag{7}$$

A schematic of Ω_k, Ω_k^{ac}, Ω_k^{si} and Ω_k^0 is given in Fig. 2.

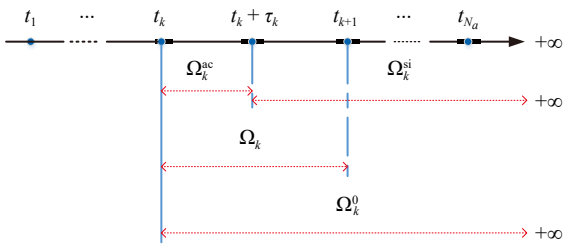


Fig. 2. Schematic diagram of Ω_k, Ω_k^{ac}, Ω_k^{si} and Ω_k^0.

D. Objective

The first aim is to propose an intermittent attack model and then analyze rigorously the stealthiness of the generated intermittent integrity attacks by the model with respect to the typical anomaly detector $\mathcal{D}$. The second objective of this paper is to design an attack detection methodology, based on a passive analytical redundancy approach (i.e., only using the system output y and the control input u in the system $\mathcal{W}$). Such a methodology is able to overcome the stealthiness of the intermittent integrity attacks and thus, can detect their occurrences.

III. Stealthiness of the Generated Intermittent Integrity Attacks

This section proposes an intermittent integrity attack model and analyzes the stealthiness (with respect to the typical anomaly detector $\mathcal{D}$) of the generated attacks. In this paper, the zero-dynamics attack model in [19] is modified to perform intermittent integrity attacks. Given the attack activating time interval Ω_k^{ac} and the silence time interval Ω_k^{si}, the attack model for the k-th attack slot, $k \in \{1, \ldots, N_a\}$, is proposed as follows:

$$\dot{\zeta}_k(t) = (A + B_a F_k)\zeta_k(t),\ \ \zeta_k(t_k) = -\Delta z_k \tag{8a}$$

$$a_k(t) = \begin{cases} F_k \zeta_k(t), & \forall\, t \in \Omega_k^{\text{ac}} \\ \left[0, a_{y,k}^T((t_k + \tau_k)^-)\right]^T, & \forall\, t \in \Omega_k^{\text{si}} \end{cases} \tag{8b}$$

$$a(t) = \sum_{i=1}^{k} a_i(t),\ \forall\, t \in \Omega_k \tag{8c}$$

where the design parameters $F_k \in \mathbb{R}^{(n_u + n_y) \times n_x}$ and $\Delta z_k \in \mathbb{R}^{n_x}$ are discussed in detail in the sequel. In addition, $a_{y,k}(t) = [0,\ I_{|K_y|}]a_k(t)$ and $a_{y,k}((t_k + \tau_k)^-) = \lim_{t \uparrow (t_k + \tau_k)} a_{y,k}(t_k + \tau_k)$. The purpose of the value $a_k(t)$ for $t \in \Omega_k^{\text{si}}$ is to guarantee the continuity of the output $y(t)$ at the pausing time instant $t_k + \tau_k$. This design of $a_k(t)$ guarantees that the generated attacks can pass through some statistical anomaly detectors (such as some detectors in [42] targeting abrupt value jumps) without being detected. The continuity issue and the corresponding design of $a_k(t)$ in (8b) arise due to the intermittent attack implementation mode, whereas typical continuous zero-dynamics attacks such as the ones in [9] and [19] do not have this issue.

The divergence of $a(t)$ generated by attack model (8) can be avoided through the intermittency performing manner described in (5). Hence, compared with the typical zero-dynamics attack in [19], the intermittent integrity attack $a(t)$ generated by attack model (8) can save the adversary's instantaneous power energy. It must be noted that the initial condition $\zeta_k(t_k) = -\Delta z_k$ in (8a), where Δz_k is not required to be zero, is more practical than the initial conditions used in [9] and [22]. Specifically, the initial condition Δz_k in (8a) can represent the difference between the true value of the system state x and the value known by the attacker. In other words, the attack model (8) does not require the knowledge of the true value of the system states, something that is implied by the nonzero Δz_k in (8a), whereas the attack models in [9] and [22] require that. In addition, as shown in the following section, similarly with [9] and [22], the intermittent integrity attacks generated by (8) are stealthy with respect to the standard anomaly detector $\mathcal{D}$ if Δz_k is sufficiently small. We assume that Δz_k satisfies the following assumption.

Assumption 1: There exist two scalars $\underline{\delta} > 0$ and $\bar{\delta} > 0$ such that the initial condition Δz_k in (8a) is bounded as follows:

$$\underline{\delta} \le |\Delta z_k| \le \bar{\delta},\ \forall\, k \in \{1, \ldots, N_a\} \tag{9}$$

where $\underline{\delta}$ and $\bar{\delta}$ are sufficiently small positive scalars that are not required to be known by the defender.

Remark 1: The lower bound $\underline{\delta}$ implies that the attacker does not need to know the true value of the system state, which is a practical assumption since the true value of the system state is hard to be precisely measured in practice due to the presence of measurement noise. The bound $\underline{\delta}$ is also not required to be known by the defender in this paper for developing the attack detection methodology. Furthermore, the (sufficiently small) upper bound $\bar{\delta}$ guarantees that the generated intermittent attacks can pass through the anomaly detector $\mathcal{D}$ without any alarms, which is analyzed in detail in the next section. In addition, the next section provides guidance on selecting a suitable $\bar{\delta}$.

In [19], it is shown that the stealthiness of zero-dynamics attacks can be violated if Δz_k is nonzero. However, [19] does not take into consideration the effects of the controller on the convergence of the system outputs in the attack scenario. In this section, both of the nonzero $\Delta z_k(t)$ and the controller $u(t)$ are considered in analyzing the stealthiness of the generated attacks by (8). To this end, a system splitting is presented. During the attack slot Ω_k, $\mathcal{W}^n$ in (2) is split into

$$\mathcal{W}_1^n : \begin{cases} \dot{x}_1^n(t) = Ax_1^n(t) \\ y_1^n(t) = Cx_1^n(t) \end{cases} \tag{10a}$$

$$\mathcal{W}_2^n : \begin{cases} \dot{x}_2^n(t) = Ax_2^n(t) + Bu^n(t) + D\omega(t) \\ y_2^n(t) = Cx_2^n(t) + v(t),\ \forall\, t \in \Omega_k \end{cases} \tag{10b}$$

where the initial conditions are $x_1^n(t_k) = 0$ and $x_2^n(t_k) = x^n(t_k)$. Such a system splitting guarantees that $x^n(t) = x_1^n(t) + x_2^n(t)$ and $y^n(t) = y_1^n(t) + y_2^n(t)$. Using a similar splitting approach, $\mathcal{W}$ in (1) is split into $\mathcal{W}_1$ and $\mathcal{W}_2$ where

$$\mathcal{W}_1 : \begin{cases} \dot{x}_1(t) = Ax_1(t) + B_a a(t) \\ y_1(t) = Cx_1(t) + D_a a(t) \end{cases} \tag{11a}$$

$$\mathcal{W}_2 : \begin{cases} \dot{x}_2(t) = Ax_2(t) + Bu(t) + D\omega(t) \\ y_2(t) = Cx_2(t) + v(t),\ \forall\, t \in \Omega_k \end{cases} \tag{11b}$$

where the initial conditions are $x_1(t_k) = -\Delta z_k$ and $x_2(t_k) = x(t_k) + \Delta z_k$. Such a splitting also guarantees that $x(t) = x_1(t) + x_2(t)$ and $y(t) = y_1(t) + y_2(t)$.

Throughout this paper, the notation Δ is used to represent the change of any variable due to an attack. For example, Δx is the change of x^n due to an attack, i.e., $\Delta x = x - x^n$. Then, the stealthiness of the intermittent integrity attacks generated by (8) is presented.

Theorem 1 (Stealthiness): Consider the weakly unobservable subspace of $\mathcal{W}_1$ in (11a) (denoted by $\mathcal{V}(\mathcal{W}_1)$), the unobservable subspace of the pair (CA, A) (denoted by $\mathcal{H}$), and the largest controlled invariant subspace of $\mathcal{W}_1$ contained in $\mathcal{H}$ (denoted by $\mathcal{V}(\mathcal{H})$). By letting $\mathcal{V}_0 = \mathcal{V}(\mathcal{W}_1) \cap \mathcal{V}(\mathcal{H})$, if the following conditions are satisfied:

$$(A + B_a F_k)\mathcal{V}_0 \subset \mathcal{V}_0,\ (C + D_a F_k)\mathcal{V}_0 = 0 \tag{12a}$$

$$\Delta z_k \in \mathcal{V}_0,\ \forall\, k \in \{1, \ldots, N_a\} \tag{12b}$$

then the change of output y of $\mathcal{W}$ in (1) due to the intermittent integrity attack $a(t)$ generated by (8) is written as

$$\Delta y(t) = \sum_{i=1}^{k} \Delta y_{2,i}(t),\ \forall\, t \in \Omega_k \tag{13}$$

where $\Delta y_{2,i}(t)$ is generated by the following system:

$$\Delta\mathcal{W}_{2,i} : \begin{cases} \Delta\dot{x}_{2,i}(t) = (A + BKC)\Delta x_{2,i}(t),\ \Delta x_{2,i}(t_i) = \Delta z_i \\ \Delta y_{2,i}(t) = C\Delta z_i(t),\ \forall\, t \in \Omega_i^0. \end{cases} \tag{14}$$

Proof: The proof is presented in Appendix A. ■

Remark 2: Comparing the space $\mathcal{V}_0$ with the controlled invariant space for generating the continuous zero-dynamics attack in [19], an additional restriction (the subspace $\mathcal{V}(\mathcal{H})$) is introduced to guarantee that Δy is continuous at the time instant $t_k + \tau_k$. Given the observable pair (C, A), $\mathcal{V}(\mathcal{H})$ is nonempty if and only if the matrix A is singular.

Remark 3: According to the incremental system (14) in Theorem 1, the output change $\Delta y_{2,k}$ due to the k-th attack converges to zero exponentially with the initial nonzero condition Δz_k, since $A + BKC$ is a Hurwitz matrix. Hence, we can conclude that the stealthiness of the generated attacks by model (8) under Assumption 1 with respect to the typical anomaly detector $\mathcal{D}$, can be guaranteed by choosing a sufficiently small initial condition Δz_k. In addition, a system with all states measurable (i.e., C is column full rank) does not possess a weakly unobservable subspace $\mathcal{V}(\mathcal{W}_1)$. Hence, no subspace $\mathcal{V}_0$ satisfying (12) exists and no stealthy intermittent integrity attack characterized by (8) exists for such a system.

Theorem 1 shows that Δz_k satisfying Assumption 1 causes a nonzero change $\Delta y_{2,k}$ at the initial time t_k, which provides an evidence that the attacks generated by (8) can be detected by some well designed passive analytical redundancy-based detection methodology. Hence, the objective of the rest of this paper is to design a passive attack detection methodology, namely by using only y and u of the system $\mathcal{W}$ in (1), for detecting the intermittent integrity attacks generated by (8) under Assumption 1.

IV. Backward-in-Time Residual Design and Analysis

In this section, an equivalent quantity of the system state change at a fixed time prior to the attack occurrence time is introduced, which is referred to as backward-in-time equivalent quantity in this paper. Also, its properties in the context of intermittent integrity attacks, are also rigorously investigated. By using the proposed backward-in-time equivalent quantity, a backward-in-time residual is designed, and the theoretical feasibility of this residual to capture the considered stealthy intermittent integrity attacks is analyzed.

At first, we suppose that the attacker selects a sufficiently large dwell time τ_k such that the k-th attack has sufficient time to cause significant damages to the system. Also, the attacker is supposed to set the time length between t_k and t_{k+1} to be sufficiently long such that sufficient energy is recovered for activating the $(k+1)$-th attack. Then, since the change $\Delta y_{2,k}$ due to the k-th attack converges to zero with an exponential speed rate (see $\Delta\mathcal{W}_{2,i}$ in Theorem 1), it is reasonable to consider that $\Delta y_{2,k}(t) \approx 0$ for $t \geq t_{k+1}$ in the design of the attack detection methodology. Therefore, based on (13) and (14), we

have $\Delta y(t) = \Delta y_2(t)$, and is described by

$$\Delta\mathcal{W}_2 : \begin{cases} \Delta\dot{z}(t) = (A + BKC)\Delta z(t),\ \Delta z(t_k) = \Delta z_k \\ \Delta y(t) = C\Delta z(t),\ \forall\, t \in \Omega_k \end{cases} \tag{15}$$

where $k \in \{1, \ldots, N_a\}$, and $\Delta z(t) = \Delta x_{2,k}(t)$ for $t \in \Omega_k$. In the rest of this paper, the detection methodology is developed based on system (15).

Intuitively, the backward-in-time equivalent quantity (mathematically defined later) is a virtual quantity of the system state change, due to the attack, at a time prior to the attack occurrence time, which is recovered from the change of the system state posterior to the attack occurrence time. Based on the backward-in-time equivalent quantity for nonlinear systems given in [36], we define a backward-in-time equivalent quantity of $\Delta z(t)$ in (15) at a time t_b, recovered based on $\Delta z(t)$ at the time t ($t \geq t_b$).

Definition 1: The backward-in-time equivalent quantity of the state $\Delta z(t)$ of the incremental system (15) at a time t_b ($t_b \leq t$), denoted by $\Delta z(t_b|t)$, is defined as

$$\Delta z(t_b|t) = \Phi(t_b, t)\Delta z(t) \tag{16}$$

where Φ is the transition matrix associated with $A + BKC$, i.e.,

$$\Phi(t_1, t_2) = e^{(A+BKC)(t_1 - t_2)},\ \forall\, t_1, t_2 \in \mathbb{R}_+. \tag{17}$$

The properties of $\Delta z(t_b|t)$ are summarized in the following lemma.

Lemma 1: Consider the attack generated by (8) satisfying Assumption 1, and the backward-in-time equivalent quantity $\Delta z(t_b|t)$ defined in (16). Then, $\Delta z(t_b|t)$ has the following properties:

1) The vector $\Delta z(t_b|t)$ is constant for $t \in \Omega_k$, i.e.,

$$\frac{d\Delta z(t_b|t)}{dt} = 0,\ \forall\, t \in \Omega_k. \tag{18}$$

2) The vector $\Delta z(t_b|t)$ is nonzero for $t \in \Omega_k$, i.e.,

$$\Delta z(t_b|t) \neq 0,\ \forall\, t \in \Omega_k. \tag{19}$$

3) (Accumulation property) For the k-th and $(k+1)$-th attack slots, their recovered backward-in-time equivalent quantities satisfy $|\Delta z(t_b|t)|_{t\in\Omega_{k+1}} \geq |\Delta z(t_b|t)|_{t\in\Omega_k}$ if Δz_{k+1} and Δz_k satisfy the following inequality:

$$|\Delta z_{k+1}| \geq \frac{\bar{\sigma}(\Phi(t_b, t_{k+1}))}{\underline{\sigma}(\Phi(t_b, t_{k+1})} |\Phi(t_{k+1}, t_k)\Delta z_k|, \quad \forall\, k \in \{1, \ldots, N_a - 1\} \tag{20}$$

where the transition matrix Φ is given in (17).

Proof: The proof is provided in Appendix B. ■

It is worth pointing out that under Assumption 1, inequality (20) is easy to satisfy. Due to the exponential convergence of the transition matrix $\Phi(t_{k+1}, t_k)$ and the sufficiently large time length $t_{k+1} - t_k$ of the k-th attack slot, $\Phi(t_{k+1}, t_k)$ is almost zero, so the right hand side of (20) is almost zero as well. Therefore, in the rest of this paper, inequality (20) is considered to hold.

Lemma 1 implies that the stealthy intermittent integrity attack can be indicated by using a residual based on $\Delta z(t_b|t)$. This is because $\Delta z(t_b|t)$ is monotonically increasing with respect to the attack slot $k \in \{1, \ldots, N_a\}$ given that (20) holds (see result 3)). Next, the backward-in-time residual is proposed by using $\Delta z(t_b|t)$ and is rigorously investigated. Based on $\Delta z(t_b|t)$, the backward-in-time equivalent quantity of $\Delta y(t)$ in (15) at the time t_b is constructed as

$$\Delta y(t_b|t) = C\Delta z(t_b|t). \tag{21}$$

Then, the backward-in-time residual, denoted by $r(t_b|t)$, is defined as follows:

$$r(t_b|t) = r(t_b) + \Delta y(t_b|t) \tag{22}$$

where the value $r(t_b)$ is given in (3). The threshold with respect to $r(t_b|t)$ is chosen as J_{th} in (4). Then, by using the residual evaluation $J(t_b|t) = |r(t_b|t)|$, the occurrence of an anomaly (fault or attack) is ascertained if at some $t > T_0$, $J(t_b|t)$ exceeds the threshold J_{th}, namely

$$J(t_b|t) > J_{th},\ \text{alarm triggering}. \tag{23}$$

Note that the backward-in-time residual $r(t_b|t)$ can indicate both faults and attacks, whereas in this paper, we mainly focus on the characteristics of $r(t_b|t)$ in terms of detecting the considered stealthy intermittent integrity attacks.

The detection methodology that integrates the residual $r(t)$ in (3) and the newly proposed backward-in-time residual $r(t_b|t)$ in (22), is referred to as ideal backward-in-time detection methodology in this paper. The terminology ideal is used since $r(t_b|t)$ is unknown due to the unknown $\Delta y(t_b|t)$. For theoretical analysis purposes at this stage, we consider $r(t_b|t)$ to be known. The ideal backward-in-time detection methodology can be found in the left hand side of Fig. 3, which includes the residual generations of $r(t)$ and $r(t_b|t)$, their evaluations, and the anomaly detection decision logic. The theoretical feasibility of $r(t_b|t)$ to detect stealthy intermittent integrity attacks is presented in the following theorem.

Theorem 2: Consider system (1), the intermittent integrity attack generated by (8) with Δz_k satisfying Assumption 1, and a fixed time instant t_b. Then, residual $r(t_b|t)$ in (22), its evaluation $J(t_b|t) = |r(t_b|t)|$, and threshold J_{th} in (4) satisfy:

1) In the absence of the intermittent integrity attack

$$J(t_b|t) \leq J_{th},\ \forall\, t \geq t_b. \tag{24}$$

2) In the presence of the intermittent stealthy integrity attack, $J(t_b|t) > J_{th}$ for all $t \in \bigcup_{i=k}^{N_a} \Omega_i$ if condition (20) holds and the k-th attack slot satisfies

$$t_k - t_b > \frac{1}{\lambda_0} \ln \frac{\underline{\sigma}(C)\underline{\delta}}{k_0 (J_{th} + |r(t_b)|)} \tag{25}$$

where $\lambda_0 > 0$ and $k_0 > 0$ are scalars satisfying $|\Phi(t_1, t_2)| \leq k_0 e^{-\lambda_0(t_1 - t_2)}$ for any $t_1 \geq t_2$.

Proof: The proof can be found in Appendix C. ■

Remark 4: Theorem 1 and Lemma 1 provide theoretical results for the ideal case that $r(t_b|t)$ is known. These theoretical findings prove rigorously that the backward-in-time residual $r(t_b|t)$ is able to trigger alarms in the presence of a stealthy intermittent integrity attack generated by (8) under Assumption 1. Result 2) in Theorem 1 is derived from result 3) in Lemma 1. Intuitively, these results imply that during an intermittent integrity attack event with various attack slots, $|\Delta z(t_b|t)|$ increases each time a new attack occurs, and maintains its new value until the next attack occurs (corresponding to (20)). As the number of intermittent attacks accumulates, at

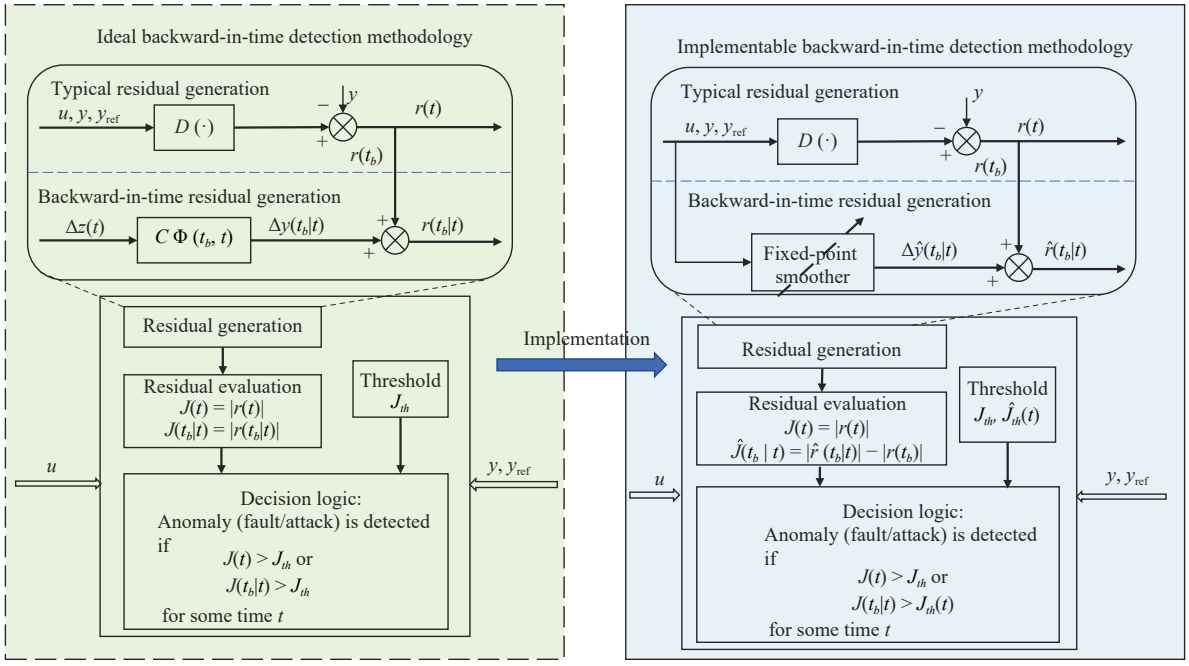


Fig. 3. Schematic diagram of the ideal and implementable backward-in-time detection methodologies.

some attack occurrence time instant, $|r(t_b|t)|$ (see (21) and (22)) exceeds the detection threshold J_{th} and thus, the intermittent integrity attack is successfully detected (corresponding to (25)).

In practice, the backward-in-time residual $r(t_b|t)$ is unknown due to the unknown $\Delta y(t_b|t)$ (see (21) and (22)). Therefore, one task of the next section is to implement a procedure for estimating optimally the unknown equivalent quantity $\Delta y(t_b|t)$ so that the residual $r(t_b|t)$ given by (22) can be implemented in practice and accordingly, a new adaptive threshold is developed based on the implemented estimation procedure.

V. Implementable Backward-in-Time Detection Methodology Design

In this section, the implementable backward-in-time detection methodology is designed. The right hand side of Fig. 3 shows the structure of the detection methodology. The optimal fixed-point smoother in Fig. 3 is first developed to estimate the unknown $\Delta y(t_b|t)$. The new backward-in-time residual $\hat{r}(t_b|t)$ and the corresponding adaptive threshold $\hat{J}_{th}(t)$ are then formulated based on the estimation results provided by the smoother. The details are given in the sequel.

A. Optimal Fixed-Point Smoother

A fixed-point smoother provides a backward-in-time estimation procedure, which produces an estimate for a signal using the past time measurements at the first stage, and then updates it using the new measurements as time progresses. Next, a fixed-point smoother in a finite time horizon $[t_b, T]$ is designed for estimating $\Delta y(t_b|t)$. Note that the fixed point t_b can be arbitrarily selected by the defender, and so it is available in the detector design.

Recalling the system splitting given in (11), in the nominal phase (i.e., $t < T_0$), we have $x(t) = z(t)$ and $y(t) = y_2(t)$ where $z(t)$ and $y_2(t)$ are the state and output of $\mathcal{W}_2$ given in (11) respectively. In addition, in the attack phase, $\Delta y(t)$ equals the output change of $\mathcal{W}_2$ (i.e., $\Delta y(t) = \Delta y_2(t)$, see Theorem 1 and $\Delta\mathcal{W}_2$ in (15)), and therefore, the smoother is designed based on the system $\mathcal{W}_2$ in (11). Let $z(t_b|t)$ denote the backward-in-time equivalent quantity of $z(t)$. Then, we have

$$z(t_b|t) = \Delta z(t_b|t) + z(t_b). \tag{26}$$

Thus, instead of estimating $\Delta z(t_b|t)$ directly, we first design a fixed-point smoother to estimate $z(t_b|t)$ and then, we use relation (26) to reconstruct $\Delta z(t_b|t)$ and relation (21) to reconstruct $\Delta y(t_b|t)$.

Now, we start by constructing the fixed-point smoother using the state augmentation approach given in [43]. To perform this task, a new state variable $\phi(t) = z(t_b|t)$ for $t \geq t_b$ is introduced, where, from result 1) in Lemma 1 and (26), $\phi(t)$ satisfies

$$\frac{d\phi(t)}{dt} = \dot{\phi}(t) = 0,\ \forall\, t \in \Omega_k. \tag{27}$$

By letting $\hat{\phi}(t)$ be the estimate of $\phi(t)$, it follows from (40) and (26) that:

$$\Delta\hat{y}(t_b|t) = C\hat{\phi}(t) - y(t_b). \tag{28}$$

Thus, based on (22), the new backward-in-time residual, denoted by $\hat{r}(t_b|t)$, is proposed as

$$\hat{r}(t_b|t) = r(t_b) + \Delta\hat{y}(t_b|t). \tag{29}$$

Motivated by the optimal residual design methodology in [29], $\hat{r}(t_b|t)$ is to be optimized for achieving robustness with respect to the disturbance ω and the measurement noise v, and sensitivity with respect to the system changes due to attacks. In the sequel, the specific objectives associated with the robustness and sensitivity are presented.

1) Robustness: The robustness considered in this paper is achieved by minimizing a piece-wise linear quadratic (LQ) cost function. Suppose that a set of smoother switching time

instants have been determined, which are given as follows:

$$t_{s,0},\ t_{s,1},\ \ldots,\ t_{s,N_s+1},\ N_s \in \mathbb{N}_+ \tag{30}$$

where $t_{s,0} = t_b$ and $t_{s,N_s+1} = T$. Correspondingly, a bank of covariance matrices[1] are introduced and given as follows:

$$\bar{\Theta}_0,\ \bar{\Theta}_1,\ \ldots,\ \bar{\Theta}_{N_s} \tag{31}$$

where $\bar{\Theta}_k \in \mathbb{R}^{2n\times 2n}$ and $\bar{\Theta}_k \geq 0$ for all $k \in \{0,\ldots,N_s\}$. The objective function associated with robustness is presented in the context of linear fractional transformation (LFT) (see LFT in [44]) in the sequel.

Consider the system $\mathcal{W}_2$ in (11b), and let $\hat{z}$ be the estimate of z, and $\hat{y}(t_b|t) = C\hat{\phi}(t)$ be the estimate of $y(t_b|t)$ in (21). Then, the smoother design is formulated to find a system $\boldsymbol{K}$ such that the LFT $\mathcal{F}_l(\boldsymbol{P},\boldsymbol{K})$ can satisfy the robustness requirement where $\mathcal{F}_l(\boldsymbol{P},\boldsymbol{K})$ is given by

$$\begin{bmatrix} \dot{q}(t) \\ y(t_b|t) - \hat{y}(t_b|t) \\ y(t) \end{bmatrix} = \boldsymbol{P}\begin{bmatrix} q(t) \\ \begin{bmatrix} \omega(t) \\ v(t) \end{bmatrix} \\ \hat{y}(t_b|t) \end{bmatrix},\ \hat{y}(t_b|t) = \boldsymbol{K}y(t). \tag{32}$$

In the above system, $q(t) = [z^T(t), \phi^T(t)]^T$, $\boldsymbol{P}$ represents the system from $[\omega^T, v^T]^T$ and $\hat{y}(t_b|t)$ to $y(t_b|t) - \hat{y}(t_b|t)$ and y, which can be written in the following matrix form:

$$\boldsymbol{P} = \left[\begin{array}{c|cc} \bar{A} & \begin{bmatrix} \bar{D} & 0 \end{bmatrix} & 0 \\ \hline \bar{L} & \begin{bmatrix} 0 & 0 \end{bmatrix} & I \\ \bar{C} & \begin{bmatrix} 0 & I \end{bmatrix} & 0 \end{array}\right]$$

with $\bar{A} = \mathrm{diag}(A, 0)$, $\bar{C} = [C,\ 0]$, $\bar{L} = [0,\ C]$ and $\bar{D} = [D^T, 0]^T$. Note that the 2-norm of a system and the 2-norm of its adjoint are equal and also, note that the adjoint of $\mathcal{F}_l(\boldsymbol{P},\boldsymbol{K})$ is $\mathcal{F}_l(\boldsymbol{P}^\sim,\boldsymbol{K}^\sim)$, in which the notation $(\cdot)^\sim$ represents the adjoint[2]. The adjoint $\mathcal{F}_l(\boldsymbol{P}^\sim,\boldsymbol{K}^\sim)$ is given as follows:

$$\begin{bmatrix} \dfrac{dp(\tau)}{d\tau} \\ \tilde{y}(t_b|\tau) \\ \tilde{\omega}(\tau) \end{bmatrix} = \boldsymbol{P}^\sim\begin{bmatrix} p(\tau) \\ \tilde{\omega}(\tau) \\ \tilde{u}(\tau) \end{bmatrix},\ \tilde{u}(\tau) = \boldsymbol{K}^\sim\tilde{\omega}(\tau) \tag{33}$$

where τ is a time-to-go variable $\tau = T + t_b - t$ with $p(t_b) = 0$, and $\boldsymbol{P}^\sim$ is given as follows:

$$\boldsymbol{P}^\sim = \left[\begin{array}{cc|cc} \bar{A}^T & & \bar{L}^T & \bar{C}^T \\ \hline \begin{bmatrix} \bar{D}^T \\ 0 \end{bmatrix} & & \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ I \end{bmatrix} \\ 0 & & I & 0 \end{array}\right].$$

Hence, the piece-wise LQ cost function is readily proposed in the context of the adjoint system (33) as follows:

$$\min_{\boldsymbol{K}^\sim}\mathcal{J} = \frac{1}{2}\sum_{k=0}^{N_s}\left\|p(\tau_{s,k})\right\|^2_{\bar{\Theta}_k} + \frac{1}{2}\sum_{k=0}^{N_s}\int_{\tau_{s,k+1}}^{\tau_{s,k}}\|\tilde{y}(t_b|\tau)\|^2_R\ d\tau \tag{34}$$

where $\tau_{s,k} = T + t_b - t_{s,k}$ for $k \in \{0,\ldots,N_s+1\}$, $R > 0$, and the weighting matrix $\bar{\Theta}_k$ is given in (31) and satisfies some structural requirements. Let P_k and Ω_k denote the covariance matrices of $z(t_{s,k}) - \hat{z}(t_{s,k})$ and $\phi(t_{s,k}) - \hat{\phi}(t_{s,k})$ respectively, and Σ_k denote the mean value of the cross term $(z(t_{s,k}) - \hat{z}(t_{s,k}))(\phi(t_{s,k}) - \hat{\phi}(t_{s,k}))^T$. Then, $P_k > 0$ and $\Omega_k > 0$, and the covariance matrix $\bar{\Theta}_k$ has the following structure:

$$\bar{\Theta}_k = \begin{bmatrix} P_k & \Sigma_k \\ \Sigma_k^T & \Omega_k \end{bmatrix},\ \forall\ k \in \{0,1,\ldots,N_s\}. \tag{35}$$

Note that by choosing $\hat{z}(t_b) = \hat{\phi}(t_b) = 0$, $z(t_b) - \hat{z}(t_b) = \phi(t_b) - \hat{\phi}(t_b)$, P_0, Σ_0 and Ω_0 satisfy

$$P_0 = \Sigma_0 = \Omega_0. \tag{36}$$

Furthermore, note also that the estimation accuracy for ϕ is higher than the one for z due to the smoothing process (see [45], [46]). Therefore, the covariance matrix of $\phi(t_{s,k}) - \hat{\phi}(t_{s,k})$ is smaller than the one of $z(t_{s,k}) - \hat{z}(t_{s,k})$, namely, P_k and Ω_k must satisfy

$$P_k \geq \Omega_k,\ \forall\ k \in \{1,\ldots,N_s\}. \tag{37}$$

2) Sensitivity: The H_- performance is introduced to quantitatively scale the sensitivity to attacks. Based on the definition of H_- in [47], the H_- performance from $z - \hat{z}$ to $\hat{y}(t_b|t)$ is defined as follows:

$$H_- = \inf_{z-\hat{z}}\frac{\int_{t_b}^T|\hat{y}(t_b|t)|^2dt}{\int_{t_b}^T|z(t) - \hat{z}(t)|^2dt},\ \forall\ z(t) - \hat{z}(t) \in \mathcal{L}_2[t_b, T]. \tag{38}$$

Thus, to guarantee the sensitivity requirement, the inequality

$$H_- \geq \alpha,\ \forall\ z(t) - \hat{z}(t) \in \mathcal{L}_2[t_b, T] \tag{39}$$

must be satisfied, where $\alpha > 0$ is a user-defined performance goal.

Remark 5: It is worth pointing out that typical optimal LQ fixed-point smoothers in [45] and [46] are obtained by minimizing the following LQ cost function given in the context of the adjoint system (33):

$$\min_{\boldsymbol{K}^\sim}\mathcal{J}_1 = \frac{1}{2}\|p(t_b)\|^2_{\bar{\Theta}_0} + \frac{1}{2}\int_{t_b}^T\|\tilde{y}(t_b|\tau)\|^2_R\ d\tau.$$

Such typical LQ smoothers possess inherently the covariance matrix wind-up problem for estimating $y(t_b|t)$. As a result, the sensitivity from $z - \hat{z}$ to $\hat{y}(t_b|t)$ vanishes as time proceeds, thereby resulting in a conflict to the H_- performance requirement in (39). More technical details in terms of this issue are given later. In this work, the fixed-point smoother, designed by minimizing the piece-wise LQ cost function (34), can reset the covariance matrix $\bar{\Theta}_k$, and is able to guarantee the sensitivity requirement (39) by designing suitable covariance matrices $\bar{\Theta}_k$ for all $k \in \{0,1,\ldots,N_s\}$, i.e., P_k, Σ_k and Ω_k in (35).

Remark 6: The adjoint system (33) facilitates the presentation of the piece-wise LQ cost function given by (34). Note that $\bar{\Theta}_k$, with the structure in (35), is a nonnegative definite matrix and nonsingular, which cannot be used for weighting the initial condition (see [48]). However, in the context of the

[1] Covariance matrix and mean value are concepts used in Kalman filtering. Since LQ optimal filters have similar form with the Kalman filter, we also use the terminologies "covariance matrix" and "mean value" for the LQ optimal filters.

[2] Regarding adjoint system of a linear system, the definition can be found in [44].

adjoint system (33), the penalty for the initial condition becomes a penalty for the terminal condition (see the term $\frac{1}{2}\sum_{k=0}^{N_s}\|p(\tau_{s,k})\|^2_{\bar{\Theta}_k}$ in (34)), and in addition, the weighting matrix is not required to be invertible. Hence, the nonsingular matrix $\bar{\Theta}_k$ in (35) can be used as the weighting matrix in the context of the adjoint system (33).

By synthesizing the objectives (34) and (39), and the requirements for the covariance matrices given in (35)−(37), a feasible way to solve the optimization problem is given in the following steps:

1) Minimize $\mathcal{J}$ for any $\bar{\Theta}_k \geq 0$;

2) Restrict P_k, Σ_k and Ω_k for all $k \in \{0,\ldots,N_s\}$ such that (36), (37) and (39) are satisfied.

In the sequel, two lemmas are rigorously derived to realize the aforementioned Steps 1) and 2) respectively. The optimal solution to (34), i.e., Step 1), is first presented.

Lemma 2: Consider linear system $\mathcal{W}_2$ in (11b) and state ϕ in (27). An optimal fixed-point smoother that minimizes (34) in the finite horizon $[t_b, T]$ is described by the following dynamics:

$$\dot{\hat{z}}(t) = A\hat{z}(t) + Bu(t) + P(t)C^TR^{-1}(y(t) - C\hat{z}(t)) \tag{40a}$$

$$\dot{\hat{\phi}}(t) = \Sigma^T(t)C^TR^{-1}(y(t) - C\hat{z}(t)) \tag{40b}$$

where $\hat{z}(t_b) = 0$ and $\hat{\phi}(t_b) = 0$. The matrices $P(t) = P^T(t) > 0$, $\Sigma(t)$ and an auxiliary matrix $\Omega(t) = \Omega^T(t) > 0$ are obtained from the solution of the following differential equations:

$$\dot{P}(t) = AP(t) + P(t)A^T - P(t)C^TR^{-1}CP(t) + DD^T \tag{41a}$$

$$\dot{\Sigma}(t) = \left(A - P(t)C^TR^{-1}C\right)\Sigma(t) \tag{41b}$$

$$\dot{\Omega}(t) = -\Sigma^T(t)C^TR^{-1}C\Sigma(t). \tag{41a}$$

In the above differential equations, at each time $t_{s,k}$ with $k \in \{0,\ldots,N_s\}$, the following switch occurs:

$$P(t_{s,k}) = P_k,\ \Sigma(t_{s,k}) = \Sigma_k,\ \Omega(t_{s,k}) = \Omega_k. \tag{42}$$

Moreover, if (37) holds, then the matrices P and Ω satisfy

$$P(t) \geq \Omega(t),\ \forall\, t \in [t_b, T]. \tag{43}$$

Proof: Consider the adjoint system (33) and the cost function $\mathcal{J}$ in (34). According to LQ control theory (see Section 5.3 in [44]), the cost function $\mathcal{J}$ is minimized by the following optimal control law $\tilde{u}(\tau)$:

$$\frac{d\hat{p}(\tau)}{d\tau} = \left(\bar{A}^T - \bar{C}^TR^{-1}\bar{C}\bar{\Theta}(\tau)\right)\hat{p}(\tau) + \bar{L}^T\tilde{\omega}(\tau),\ \hat{p}(t_b) = 0$$

$$\tilde{u}(\tau) = -R^{-1}\bar{C}\bar{\Theta}(\tau)\hat{p}(\tau)$$

where the covariance matrix $\bar{\Theta}(\tau)$ is generated by

$$-\frac{d\bar{\Theta}(\tau)}{d\tau} = \bar{A}\bar{\Theta}(\tau) + \bar{\Theta}(\tau)\bar{A}^T - \bar{\Theta}(\tau)\bar{C}^TR^{-1}\bar{C}\bar{\Theta}(\tau) + \bar{D}\bar{D}^T$$

with the following switches:

$$\bar{\Theta}(\tau_{s,k}) = \bar{\Theta}_k,\ \forall\, k = \{0,1,\ldots,N_s\}.$$

Then, the optimal smoother is obtained as the adjoint system of $\tilde{u}(\tau)$ and is given as follows:

$$\dot{\hat{q}}(t) = (\bar{A} - \bar{\Theta}(t)\bar{C}^TR^{-1}\bar{C})\hat{q}(t) + \bar{\Theta}(t)\bar{C}^TR^{-1}y(t) \tag{44a}$$

$$\hat{y}(t_b|t) = \bar{L}\hat{q}(t) \tag{44b}$$

where $\hat{q} = [\hat{z}^T, \hat{\phi}^T]^T$ and $\hat{q}(t_b) = 0$, and the covariance matrix $\bar{\Theta}(\tau)$ is generated by

$$\dot{\bar{\Theta}}(t) = \bar{A}\bar{\Theta}(t) + \bar{\Theta}(t)\bar{A}^T - \bar{\Theta}(t)\bar{C}^TR^{-1}\bar{C}\bar{\Theta}(t) + \bar{D}\bar{D}^T \tag{45}$$

with the following switches:

$$\bar{\Theta}(t_{s,k}) = \bar{\Theta}_k,\ \forall\, k \in \{0,1,\ldots,N_s\}. \tag{46}$$

Thus, from (44), the smoother (40) can be obtained.

We now proceed to derive the differential equations given in (41) and switches in (42). By letting

$$\bar{\Theta}(t) = \begin{bmatrix} P(t) & \Sigma(t) \\ \Sigma^T(t) & \Omega(t) \end{bmatrix}.$$

It follows from (45) that:

$$\begin{aligned}
\begin{bmatrix} \dot{P} & \dot{\Sigma} \\ \dot{\Sigma}^T & \dot{\Omega} \end{bmatrix} &= \begin{bmatrix} A - P(t)C^TR^{-1}C & 0 \\ -\Sigma^T(t)C^TR^{-1}C & 0 \end{bmatrix}\begin{bmatrix} P & \Sigma \\ \Sigma^T & \Omega \end{bmatrix} \\
&\quad + \begin{bmatrix} P & \Sigma \\ \Sigma^T & \Omega \end{bmatrix}\begin{bmatrix} A^T - C^TR^{-1}CP(t) & -C^TR^{-1}C\Sigma(t) \\ 0 & 0 \end{bmatrix} \\
&\quad + \begin{bmatrix} D & -P(t)C^TR^{-1} \\ 0 & -\Sigma^T(t)C^TR^{-1} \end{bmatrix}\begin{bmatrix} I & \\ & R \end{bmatrix}\begin{bmatrix} D^T & -R^{-1}CP(t) \\ 0 & -R^{-1}C\Sigma(t) \end{bmatrix}.
\end{aligned}$$

By simplifying the above differential Riccati equation, the differential equations (41) can be obtained. In addition, from (35) and (46), the switches in (42) can also be obtained.

Regarding the result (43), by letting $Y(t) = P(t) - \Omega(t)$ and $A_Y = A - \Omega C^TR^{-1}C - \frac{1}{2}YC^TR^{-1}C$, it follows from (41a) and (41c) that:

$$\begin{aligned}
\dot{Y} &= AY + YA^T + A\Omega + \Omega A^T - YC^TR^{-1}CY - \Omega C^TR^{-1}CY \\
&\quad - YC^TR^{-1}C\Omega - YC^TR^{-1}C\Omega + \Omega C^TR^{-1}C\Omega + DD^T \\
&\quad + \Sigma^T(t)C^TR^{-1}C\Sigma(t) \\
&= A_YY + YA_Y^T + DD^T + \Sigma^T(t)C^TR^{-1}C\Sigma(t) \\
&\quad + \begin{bmatrix} I & \Omega \end{bmatrix}\begin{bmatrix} 0 & A \\ A^T & C^TR^{-1}C \end{bmatrix}\begin{bmatrix} I \\ \Omega \end{bmatrix}.
\end{aligned}$$

The above equation indicates that

$$\dot{Y}(t) \geq A_YY(t) + YA_Y^T(t),\ \forall\, t \in [t_{s,k}, t_{s,k+1}),\ \forall\, k \in \{0,\ldots,N_s\}.$$

Under the condition (37), we can obtain that $Y(t_{s,k}) \geq 0$. Therefore, based on Theorem 4.1.2 in [49], $Y(t) \geq 0$ for $t \in [t_{s,k}, t_{s,k+1})$ and $k \in \{0,\ldots,N_s\}$ can be obtained and the result (43) follows. ■

It can be observed from (40b) and (41b) that in the case of a typical LQ smoother (obtained by minimizing the cost function $\mathcal{J}_1$ and without matrix resetting (42)), the sensitivity from $\hat{y}(t_b|t)$ to $z - \hat{z}$, characterized by the matrix $\Sigma(t)$, decreases as time proceeds. The reason for this is that the solution $\Sigma(t)$ of the differential equation (41b) converges to zero as the time progresses, which is the aforementioned covariance matrix wind-up problem. The matrix resetting at the time instant t_k characterized in (42) provides an alternative way to maintain

the sensitivity. In the following lemma, feasible resetting matrices P_k, Σ_k and Ω_k that satisfy (36), (37) and (39) are presented.

Lemma 3: Consider the piece-wise fixed-point smoother given in Lemma 2 and the switching time instants given in (30). Consider also the differential equations (41) without the switches (42). The requirements (36) and (37) are guaranteed if

$$P_k = P(t_{s,k}),\ \Sigma_k = \Sigma(t_{s,k}) = \Theta_k,\ \Omega_k = \Omega(t_{s,k}),\ \forall\, k \in \{0,\ldots,N_s\} \tag{47}$$

where $P(t_{s,k})$, $\Sigma(t_{s,k})$ and $\Omega(t_{s,k})$ are the values at $t_{s,k}$ of the solutions to the differential equations (41) without the switches (42) and under the following initial conditions:

$$P(t_b) = \Sigma(t_b) = \Omega(t_b) = \Theta_k,\ \forall\, k \in \{0,\ldots,N_s\}. \tag{48}$$

Moreover, the H_- performance index requirement in (39) is guaranteed if Θ_k in (48) satisfies

$$\underline{\sigma}(\Theta_k) \geq \frac{\alpha}{\underline{\sigma}(CR^{-1}C^T)\underline{\sigma}(C)\underline{\sigma}\left(e^{(A-P(t_{s,k+1})C^TR^{-1}C)(t_{s,k+1}-t_b)}\right)},$$
$$\forall\, k \in \{0,1,\ldots,N_s\}. \tag{49}$$

Proof: According to (47), P_k and Ω_k are the matrices associated with a typical LQ fixed-point smoother without the matrix resetting. Hence, (37) can be guaranteed directly.

According to (40b), we can obtain that

$$\int_{t_b}^{T} |\hat{y}(t_b|t)|^2 dt \geq \underline{\sigma}^2(C\Sigma^T(t)C^TR^{-1}C)|z(t)-\hat{z}(t)|^2$$

which indicates that the H_- performance requirement in (39) can be guaranteed if

$$\underline{\sigma}^2(C\Sigma^T(t)C^TR^{-1}C) \geq \alpha^2,\ \forall t \in [t_b,T].$$

Based on the inequality $\underline{\sigma}(C\Sigma^T(t)C^TR^{-1}C) \geq \underline{\sigma}(C^TR^{-1}C)\times \underline{\sigma}(\Sigma(t))\underline{\sigma}(C)$, a sufficient condition to guarantee the above inequality is obtained as

$$\underline{\sigma}^2(\Sigma(t)) \geq \frac{\alpha^2}{\underline{\sigma}^2(C^TR^{-1}C)\underline{\sigma}^2(C)},\ \forall\, t \in [t_b,T]. \tag{50}$$

Let $X(t) = \Sigma(t)\Sigma^T(t)$ where $X(t_b) = \Theta_k\Theta_k^T$. Then, it follows from (41b) without the switches (42) that:

$$\begin{aligned}\dot{X}(t) &= \dot{\Sigma}(t)\Sigma^T(t) + \Sigma(t)\dot{\Sigma}^T(t)\\ &= \left(A - P(t)C^TR^{-1}C\right)X(t) + X(t)\left(A - P(t)C^TR^{-1}C\right)^T\end{aligned}$$

where the system $\dot{x} = (A - P(t)C^TR^{-1}C)x$ is exponentially stable. Moreover, $X(t)$ can be written as

$$X(t) = e^{\left(A-P(t)C^TR^{-1}C\right)^T(t-t_b)}X(t_b)e^{\left(A-P(t)C^TR^{-1}C\right)(t-t_b)} \tag{51}$$

which indicates that $X(t)$ is monotonically decreasing with respect to time and $X(t) \leq X(t_b) = \Theta_k\Theta_k^T$ for $t \geq t_b$. Thus, the inequality (50) can be guaranteed, if at the end of k-th time interval (i.e., $t_{s,k+1}$), the following condition is satisfied:

$$\underline{\sigma}(X(t_{s,k+1})) \geq \frac{\alpha^2}{\underline{\sigma}^2(C^TR^{-1}C)\underline{\sigma}^2(C)},\ \forall\, k \in \{0,1,\ldots,N_s\}.$$

It then follows from (51) and $X(t_b) = \Theta_k\Theta_k^T$ that the above inequality can be guaranteed by (49). ■

Subsequently, by synthesizing the results in Lemmas 2 and 3, a feasible solution to minimize $\mathcal{J}$ in (34) and to satisfy the restrictions (36), (37) and (39) is presented by the following theorem.

Theorem 3: Consider linear system $\mathcal{W}_2$ in (11b) and state ϕ in (27). The fixed-point smoother (40) with $P(t)$, $\Sigma(t)$ and $\Omega(t)$ determined by (41) and the matrix resetting given by (42), minimizes the cost function $\mathcal{J}$ in (34) in the finite time horizon $[t_b,T]$. Moreover, by constructing P_k, Σ_k and Ω_k as in (47) and (48), and by choosing Θ_k to satisfy (49) for $k \in \{0, 1,\ldots,N_s\}$, the H_- performance requirement in (39) and the requirements (36) and (37) are guaranteed simultaneously.

B. Residual Evaluation and Threshold Generation

In this section, the detection residual $\hat{r}(t_b|t)$ in (29) is evaluated and an adaptive threshold is generated. We start by evaluating $\hat{r}(t_b|t)$. Let $e_z(t) = z(t) - \hat{z}(t)$, $e_\phi(t) = \phi(t) - \hat{\phi}(t)$ and $e_y(t) = y(t_b|t) - \hat{y}(t_b|t)$ be the estimation errors of x, ϕ and $y(t_b|t)$, respectively. Then, from (11b), (27) and (40), the error system is obtained as follows:

$$\dot{e}_z(t) = (A - P(t)C^TR^{-1}C)e_z(t) + D\omega(t) - P(t)C^TR^{-1}v(t) \tag{52a}$$

$$\dot{e}_\phi(t) = -\Sigma^T(t)C^TR^{-1}Ce_z(t) - \Sigma^T(t)C^TR^{-1}v(t) \tag{52b}$$

$$e_y(t) = Ce_\phi(t) \tag{52c}$$

where the initial conditions are $e_z(t_b) = z(t_b)$ and $e_\phi(t_b) = \phi(t_b)$. In addition, from $\hat{y}(t_b|t) = y(t_b|t) - e_y(t)$, $\Delta\hat{y}(t_b|t) = \hat{y}(t_b|t) - y(t_b)$ and $\Delta y(t_b|t) = y(t_b|t) - y(t_b)$, $\hat{r}(t_b|t)$ in (29) can be split into

$$\hat{r}(t_b|t) = r(t_b) + \Delta y(t_b|t) - e_y(t). \tag{53}$$

As in the case of fault diagnosis literature, the residual evaluation should ensure that in the non-attack case, the value of the evaluation function is close or equal to zero under ideal conditions (e.g., no disturbance, no noise and no modeling uncertainty). Thus, from (53), the evaluation function is proposed as follows:

$$\hat{J}(t_b|t) = |\hat{r}(t_b|t)| - |r(t_b)| \tag{54}$$

where the correction term $-|r(t_b)|$ guarantees that $\hat{J}(t_b|t)$ is close to zero in the absence of attacks.

Next, a bound $\hat{J}_{th}(t)$ (adaptive threshold) of the residual evaluation function $\hat{J}(t_b|t)$ is derived. To this end, a lemma required to generate the threshold is given.

Lemma 4: Let $\Phi_k(t,\tau)$ be the transition matrix associated with the matrix $A - P(t)C^TR^{-1}C$ for $t,\tau \in [t_{s,k},t_{s,k+1})$. Then, there exist known scalars $\beta_k > 0$ and $\lambda_k > 0$ such that

$$|\Phi_k(t,\tau)| \leq \beta_k e^{-\lambda_k(t-\tau)},\ \forall\, t,\tau \in [t_{s,k},t_{s,k+1}). \tag{55}$$

Proof: Note that since the pair (A,D) is stabilizable and (C,A) is observable, the system $\dot{x} = (A - P(t)C^TR^{-1}C)x$ is exponentially stable. Thus, $x(t) = \Phi_k(t,\tau)x(\tau)$ converges exponentially to zero during the time interval $t \in [t_{s,k},t_{s,k+1})$. Hence, such scalars β_k and λ_k that satisfy (55) exist. ■

In addition, the disturbance $\omega(t)$ and the measurement noise $v(t)$ are supposed to satisfy the following assumption.

Assumption 2: The disturbance $\omega(t)$ and the measurement

noise $v(t)$ are bounded by $\bar{\omega} > 0$ and $\bar{v} > 0$, respectively, i.e.,

$$|\omega(t)| \le \bar{\omega},\ |v(t)| \le \bar{v},\ \forall\, t \in \mathbb{R}_+ \tag{56}$$

where $\bar{\omega}$ and $\bar{v}$ are known scalars by the defender.

Remark 7: Such an assumption is commonly used in anomaly diagnosis literature (see, e.g., [29], [32], [40]) for guaranteeing robustness and avoiding false alarms. The bound $\bar{\omega}$ can be obtained a priori by the defender by some experimental tests to the considered CPS, while $\bar{v}$ can be obtained by exploiting a priori knowledge of the sensor bias deviation based on the technical characteristics of the sensors.

In the following theorem, the adaptive threshold is presented.

Theorem 4 (Robustness): Consider closed-loop CPS $\mathcal{W}$ in (1) and anomaly detector described in (3) and (4). Also, consider the piece-wise fixed-point smoother (40) described in Theorem 3, the residual (29) and the residual evaluation function (54). Under Assumptions 1 and 2, and in the non-attack case (no intermittent integrity attacks), the residual evaluation function $\hat{J}(t_b|t)$ in (54) is bounded by the adaptive threshold $\hat{J}_{th}(t)$ as follows:

$$\hat{J}(t_b|t) \le \hat{J}_{th}(t),\ \forall t \in [t_{s,k}, t_{s,k+1}) \tag{57}$$

where $\hat{J}_{th}(t) = |C|\varepsilon_{\phi,k}(t)$ with

$$\varepsilon_{\phi,k}(t) = \varepsilon_{\phi,k-1}(t_{s,k}^-) + |\Sigma_k| \cdot |C^T R^{-1}| \beta_k \int_{t_{s,k}}^{t} e^{-\lambda_k(\tau - t_{s,k})} (\varepsilon_{z,k}(\tau) + \bar{v})\, d\tau \tag{58}$$

$$\varepsilon_{z,k}(t) = \beta_k e^{-\lambda_k(t-t_{s,k})} |e_z(t_{s,k}^-)| + \frac{\beta_k \left(|D|\bar{\omega} + |P_k| \cdot |C^T R^{-1}| \bar{v}\right)}{\lambda_k} \left(1 - e^{\lambda_k(t_{s,k} - t)}\right). \tag{59}$$

In the above equations, β_k and λ_k are specified in Lemma 4.

Proof: Based on (53) and by using the triangle inequality, the evaluation $\hat{J}(t_b|t)$ in (54) satisfies

$$\hat{J}(t_b|t) \le |e_y(t)| + |\Delta y(t_b|t)|. \tag{60}$$

Note that in the non-attack case, $\Delta z(t_b|t) = 0$ and $\Delta y(t_b|t) = 0$. It then follows from (60) that the threshold $\hat{J}_{th}(t)$ is chosen as:

$$\hat{J}_{th}(t) = \sup_{\Delta y(t_b|t)=0} \hat{J}(t_b|t) = \sup_{\Delta y(t_b|t)=0} |e_y(t)|.$$

By solving the differential equation (52a), $e_z(t)$ can be explicitly expressed as

$$e_z(t) = \Phi_k(t, t_{s,k}) e_z(t_{s,k}^-) + \int_{t_{s,k}}^{t} \Phi_k(t,\tau) \left(D\omega(\tau) - P(\tau) C^T R^{-1} v(\tau)\right) d\tau, \quad \forall t \in [t_{s,k}, t_{s,k+1}).$$

Note that based on Theorem 3.1.1 in [44], $P(t)$ in (41a) is monotonically decreasing with respect to t and thus, $P(t) \le P(t_{s,k}) = P_k$ for $t \in [t_{s,k}, t_{s,k+1})$. Thus, based on Assumption 2 and Lemma 4, we can obtain

$$\begin{aligned} |e_z(t)| &\le \beta_k e^{-\lambda_k(t-t_{s,k})} |e_z(t_{s,k}^-)| + \int_{t_{s,k}}^{t} \beta_k e^{-\lambda_k(t-\tau)} \left(|D|\bar{\omega} + |P_k| \times |C^T R^{-1}| \bar{v}\right) d\tau \\ &= \beta_k e^{-\lambda_k(t-t_{s,k})} |e_z(t_{s,k}^-)| + \frac{\beta_k \left(|D|\bar{\omega} + |P_k| \times |C^T R^{-1}| \bar{v}\right)}{\lambda_k} \left(1 - e^{\lambda_k(t_{s,k} - t)}\right). \end{aligned}$$

Thus, $\varepsilon_{z,k}(\cdot)$ in (59) is obtained. In addition, by solving the differential equation (52b), we can obtain

$$e_\phi(t) = e_\phi(t_{s,k}^-) - \int_{t_{s,k}}^{t} \Sigma^T(\tau) C^T R^{-1} (e_z(\tau) + v(\tau))\, d\tau,\ \forall\, t \in [t_{s,k}, t_{s,k+1}).$$

Note that it follows from (51) that $\Sigma(t)$ satisfies:

$$\begin{aligned} |\Sigma(t)| &= \sqrt{|X(t)|} = \sqrt{|X(t_{s,k})|} \times |\Phi_k(t, t_{s,k})| \\ &\le |\Sigma_k| \beta_k e^{-\lambda_k(t-t_{s,k})},\ \forall\, t \in [t_{s,k}, t_{s,k+1}). \end{aligned}$$

Thus, from $|e_z(t)| \le \varepsilon_{z,k}(\cdot)$ and $|v(t)| \le \bar{v}$ in Assumption 2, and by using the triangle inequality, we can obtain

$$|e_\phi(t)| \le |e_\phi(t_{s,k}^-)| + |\Sigma_k| \beta_k \int_{t_{s,k}}^{t} e^{-\lambda_k(\tau - t_{s,k})} |C^T R^{-1}| (\varepsilon_{z,k}(\tau) + \bar{v})\, d\tau.$$

Since $|e_\phi(t_{s,k}^-)| \le \varepsilon_{\phi,k-1}(t_{s,k}^-)$, then $\varepsilon_{\phi,k}(\cdot)$ in (58) is obtained. Hence, from $|e_y| \le |C| \cdot |e_\phi(t)|$, (57) is obtained. ■

Remark 8: The threshold $\hat{J}_{th}(t)$ in (57) is calculated iteratively since $\varepsilon_{z,k}$ and $\varepsilon_{\phi,k}$ for the time interval $[t_{s,k}, t_{s,k+1})$ rely on $e_z(t_{s,k}^-)$ and $\varepsilon_{\phi,k-1}(t_{s,k}^-)$ respectively from the previous time interval $[t_{s,k-1}, t_{s,k})$. In addition, $\varepsilon_{z,k}$ in (59) cannot be used directly since $|e_z(t_{s,k}^-)|$ is not available to the defender. To overcome this, it is reasonable to suppose that there exists a scalar $\delta_0 > 0$ such that $|e_z(t_{s,k})| \le \delta_0$. Such δ_0 does not affect significantly the final detection result since the term $\beta_k e^{-\lambda_k(t-t_{s,k})} |e_z(t_{s,k}^-)| \le \beta_k e^{-\lambda_k(t-t_{s,k})} \delta_0$ converges to zero exponentially. Therefore, the designer can select a sufficiently large δ_0 for implementing the threshold $\hat{J}_{th}(t)$.

Based on the residual $\hat{r}(t_b|t)$ in (29), the evaluation function $\hat{J}(t_b|t)$ in (54), and the threshold $\hat{J}_{th}(t)$ in (57), the attack occurrence decision principle is given as follows: if there exists a time $t > t_b$ such that $\hat{J}(t_b|t)$ in (54) exceeds the threshold $\hat{J}_{th}(t)$, i.e., $\hat{J}(t_b|t) > \hat{J}_{th}(t)$, then an alarm is triggered to indicate the presence of an attack. The detection time T_d of the attack is defined as the first time instant when the inequality $\hat{J}(t_b|t) > \hat{J}_{th}(t)$ holds for a given t_b, i.e.,

$$T_d(t_b) = \inf \left\{ t > t_b \,\middle|\, \hat{J}(t_b|t) > \hat{J}_{th}(t) \right\}. \tag{61}$$

In addition, as Fig. 3 shows, by combining the residuals $r(t)$ and $\hat{r}(t_b|t)$, the evaluation functions $J(t)$ and $\hat{J}(t_b|t)$, and the thresholds J_{th} and $\hat{J}_{th}(t)$, the occurrence of an anomaly (fault or attack) is decided if there exists a time $t > t_b$ such that $J(t) > J_{th}$ or $\hat{J}(t_b|t) > \hat{J}_{th}(t)$.

Algorithm 1 provides in concise form the steps required for implementing the smoother, generating the backward-in-time residual $\hat{r}(t_b|t)$ and threshold $\hat{J}_{th}(t)$, and the decision principle for detecting the considered stealthy intermittent integrity attacks.

Algorithm 1 Backward-in-Time Attack Detection Algorithm

procedure SMOOTHER(t_b, T, $t_{s,k}$, α, R) // Theorem 3;
  $k \leftarrow 0$;
  repeat // Lemma 3;
    Choose Θ_k based on (49);
    $P(t_b) \leftarrow \Theta_k$, $\Sigma(t_b) \leftarrow \Theta_k$, $\Omega(t_b) \leftarrow \Theta_k$; // (48);
    $P_k \leftarrow P(t_{s,k})$, $\Sigma_k \leftarrow \Sigma(t_{s,k})$, $\Omega_k \leftarrow \Omega(t_{s,k})$; //solve the differential equations in (41); **return** P_k, Σ_k, Ω_k;
    $k \leftarrow k+1$;
  until $k = N_s$
  $P(t_{s,0}) \leftarrow \Theta_0$, $\Sigma(t_{s,0}) \leftarrow \Theta_0$, $\Omega(t_{s,0}) \leftarrow \Theta_0$; // Lemma 2;
  Solve differential equations in (41) for $t \in [t_{s,0}, t_{s,1})$;
  return $P(t)$, $\Sigma(t)$, $\Omega(t)$ for $t \in [t_{s,0}, t_{s,1})$;
  $k \leftarrow 1$;
  repeat
    $P(t_{s,k}) \leftarrow P_k$, $\Sigma(t_{s,k}) \leftarrow \Sigma_k$, $\Omega(t_{s,k}) \leftarrow \Omega_k$; // (42);
    Solve differential equations in (41) for $t \in [t_{s,k}, t_{s,k+1})$;
    return $P(t)$, $\Sigma(t)$, $\Omega(t)$ for $t \in [t_{s,k}, t_{s,k+1})$;
    $k \leftarrow k+1$;
  until $k = N_s$
  Construct the smoother as follows: // (40);

$$\dot{\hat{z}}(t) = A\hat{z}(t) + Bu(t) + P(t)C^T R^{-1}(y(t) - C\hat{z}(t))$$

$$\dot{\hat{\phi}}(t) = \Sigma^T(t)C^T R^{-1}(y(t) - C\hat{z}(t))$$

end procedure
//
procedure RESIDUAL($r(t_b), y(t_b), \hat{\phi}(t)$)
  Residual $\hat{r}(t_b|t) = r(t_b) + C\hat{\phi}(t) - y(t_b)$; // (29) and $\Delta\hat{y}(t_b|t) = C\hat{\phi}(t) - y(t_b)$;
  Evaluation $\hat{J}(t_b|t) = |\hat{r}(t_b|t)| - |r(t_b)|$; // (54);
end procedure
//
procedure THRESHOLD($P_k, \Sigma_k, \bar{\omega}, \bar{v}$) // Theorem 4;
  $k \leftarrow 0$
  repeat
    Determine β_k, λ_k satisfying (55); // Lemma 4;
    Determine δ_0 satisfying $|e_z(t_{s,k})| \le \delta_0$;
    Calculate $\varepsilon_{z,k}(\cdot)$, $\varepsilon_{\phi,k}(\cdot)$ for $t \in [t_{s,k}, t_{s,k+1})$; // (59) and (58);
  until $k = N_s$
  Threshold $\hat{J}_{th}(t) = |C|\varepsilon_{\phi,k}(t)$ for $t \in [t_{s,k}, t_{s,k+1})$; // (57);
end procedure
//
procedure DECISION PRINCIPLE($\hat{J}(t_b|t), \hat{J}_{th}(t)$)
  if $\hat{J}(t_b|t) > \hat{J}_{th}(t)$ **then** an alarm is triggered;
  else no attack is detected;
  end if
end procedure

VI. Attack Detectability Analysis

In this section, the attack detectability of the developed backward-in-time detection methodology characterized by the residual $\hat{r}(t_b|t)$ in (29), the evaluation function $\hat{J}(t_b|t)$ in (54), and the threshold $\hat{J}_{th}(t)$ in (57), is investigated rigorously, characterizing quantitatively the class of detectable intermittent integrity attacks.

Theorem 5 (Detectability): Consider closed-loop CPS $\mathcal{W}$ in (1) and anomaly detector described in (3) and (4). The attack detection decision scheme, characterized by the piece-wise fixed-point smoother (40) described in Theorem 3, the residual (29), residual evaluation function (54) and detection threshold (57), guarantees that an intermittent integrity attack generated by the model (8) can be detected at a time $T_d \ge T_0 > t_b$, i.e., $\hat{J}(t_b|T_d) > \hat{J}_{th}(T_d)$, if for the given fixed time instant t_b, there exists an attack slot k such that

$$t_k - t_b > +\frac{1}{\lambda_0}\ln\frac{\underline{\sigma}(C)\underline{\delta}}{k_0\left(J_{th} + 2\hat{J}_{th}(T_d) + |r(t_b)|\right)} \tag{62}$$

where $\underline{\delta}$ is given in Assumption 1, k_0 and λ_0 are specified in Theorem 2, J_{th} is given in (4) and $\hat{J}_{th}$ is given in (57).

Proof: For $\hat{r}(t_b|t) = r(t_b|t) - e_y(t)$ given in (53), by using the reverse triangle inequality, we can obtain

$$|\hat{r}(t_b|T_d)| \ge |r(t_b|T_d)| - |e_y(T_d)|. \tag{63}$$

From (54) and $|r(t_b)| < J_{th}$, to detect an attack at the time instant T_d, i.e., $\hat{J}(t_b|T_d) > \hat{J}_{th}(T_d)$, the following inequality must hold:

$$|\hat{r}(t_b|T_d)| > J_{th} + \hat{J}_{th}(T_d). \tag{64}$$

Then, from (63) and the fact that $|e_y(t)| \le \hat{J}_{th}(t)$ for $t \le T_d$, a sufficient condition to guarantee (64) can be obtained as follows:

$$|r(t_b|T_d)| > J_{th} + 2\hat{J}_{th}(T_d). \tag{65}$$

By using the same reasoning logic with the proof of Theorem 2, we can obtain that for the fixed time instant t_b, if there exists an attack slot k satisfying (62), then the inequality (65) can be guaranteed for any $T_d \in \bigcup_{i=k}^{N_a}\Omega_i$. Hence, the result is proved. ■

Theorem 5 is a theoretical result that cannot be checked *a priori*. It is important to note that according to Theorem 5, the attack detection by the developed implementable backward-in-time detection methodology, characterized by the residual $\hat{r}(t_b|t)$ in (29), the evaluation function $\hat{J}(t_b|t)$ in (54), and the threshold $\hat{J}_{th}(t)$ in (57), is guaranteed by selecting a small t_b (i.e., close to zero) so that condition (62) can hold. In other words, as time progresses and at each attack activation time t_k, the left side of (62) increases and at some time, will exceed the right hand side of (62), leading to the detection of the intermittent integrity attack.

VII. Simulation

In this section, a numerical simulation example based on a linear time-invariant system in the form of system (1) is presented. The system matrices are given as follows:

$$A = \begin{bmatrix} -3.25 & 1 & 0 \\ 1 & -3 & 0 \\ 0 & 0 & 0 \end{bmatrix},\ B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$C = [0 \quad 1 \quad 1],\ D = \begin{bmatrix} 1.0000 & 0 \\ 0 & 0 \\ 1.6667 & 1.6667 \end{bmatrix}$$

where the pair (C, A) is observable and (A, D) is stabilizable. The output control gain is given by $K = -4.7333$ and the reference signal $y_{\text{ref}} = 0$. In addition, for the simulation purpose, the disturbance $\omega(t)$ is given by

$$\omega(t) = [0.2\sin(5t), 0.1\sin(3t)]^T$$

and the measurement noise $v(t)$ is chosen at each time instant as a uniformly distributed random number ranging from -0.3 to 0.3. The anomaly detector $\mathcal{D}$ in (3) is designed based on [29]. The residual is designed to satisfy the optimal $\mathcal{H}_\infty$ performance, i.e., $\int_0^{15} r^T(t)r(t)dt \le 2\int_0^{15}[\omega^T(t), v^T(t)]^T[\omega^T(t), v^T(t)]dt$ and the threshold is chosen as $J_{th} = 5$ based on the fault and attack free operation.

A. Intermittent Integrity Attack

In this part, the intermittent integrity attack used in this simulation is given. The attacker is supposed to know the system matrices A, B and C, and is able to compromise all the sensors and actuators, i.e., $\Gamma_u = 1$ and $\Gamma_y = 1$. The attack activating time instants t_k are given first as follows:

$$t_1 = 1\,\text{s},\ t_2 = 3\,\text{s},\ t_3 = 5\,\text{s},\ t_4 = 7\,\text{s},\ t_5 = 9\,\text{s},\ t_6 = 11\,\text{s},\ t_7 = 15\,\text{s}$$

and the same dwell time is used for all attacks, i.e., $\tau_k = 1$ s for all $k \in \{1, \ldots, 6\}$. In the sequel, the design parameters of the attack model (8) are calculated. Based on Theorem 1, and by using the geometric approach toolbox in [50], we can obtain that the subspace $\mathcal{V}_0$ satisfying Theorem 1 is $\mathcal{V}_0 = [0, 0, -1]^T$ and further, a feasible F_k satisfying (12a) is calculated as

$$F_k = \begin{bmatrix} 0 & 0 & 1.200 \\ 0 & 0 & -1.000 \end{bmatrix},\ \forall\, k \in \{1, \ldots, 6\}.$$

According to the obtained $\mathcal{V}_0$, Δz_k satisfying (12b) is chosen as

$$\Delta z_1 = [0, 0, -0.2091],\ \Delta z_2 = [0, 0, -0.2210]$$

$$\Delta z_3 = [0, 0, -0.1191],\ \Delta z_4 = [0, 0, -0.1531]$$

$$\Delta z_5 = [0, 0, -0.1531],\ \Delta z_6 = [0, 0, -0.1095].$$

Thus, the design parameters of the attack model (8) have been selected and the attack signals for the attack activating time interval Ω_k^{ac} can be generated. Next, the attack signals during the attack silence time interval Ω_k^{si}, i.e., $a_{y,k}((t_k + \tau_k)^-)$, are given based on (8b) as follows:

$$a_1(t) = [0, -1.2919]^T,\ k = 1,\ \forall t \in [2\,\text{s}, 15\,\text{s})$$

$$a_2(t) = [0, -1.2392]^T,\ k = 2,\ \forall t \in [4\,\text{s}, 15\,\text{s})$$

$$a_3(t) = [0, -1.2060]^T,\ k = 3,\ \forall t \in [6\,\text{s}, 15\,\text{s})$$

$$a_4(t) = [0, -1.0931]^T,\ k = 4,\ \forall t \in [8\,\text{s}, 15\,\text{s})$$

$$a_5(t) = [0, -1.0931]^T,\ k = 5,\ \forall t \in [10\,\text{s}, 15\,\text{s})$$

$$a_6(t) = [0, -1.2432]^T,\ k = 6,\ \forall t \in [12\,\text{s}, 15\,\text{s}).$$

Thus, the intermittent integrity attack signal for the attack activating time interval Ω_k^{ac} and the attack silence time interval Ω_k^{si} is generated.

The attack signal and its effects on the system are shown in Figs. 4–6, respectively. As it is shown in Fig. 4, the attack signal $a_y(t)$ is continuous at the attack pausing time instants $\{2\,\text{s}, 4\,\text{s}, 6\,\text{s}, 8\,\text{s}, 10\,\text{s}, 12\,\text{s}\}$. Moreover, the resulting system output y and its change in Fig. 6 have no jump (abrupt change) at each of these attack pausing time instants. By comparing the system states x in the attack case with the state x^n in the nominal case in Fig. 5, we can observe that the system state x_3 is significantly affected by the injected intermittent integrity attacks. However, in contrast to y^n in the nominal case, the system output y in Fig. 6 has a very small increment Δy at

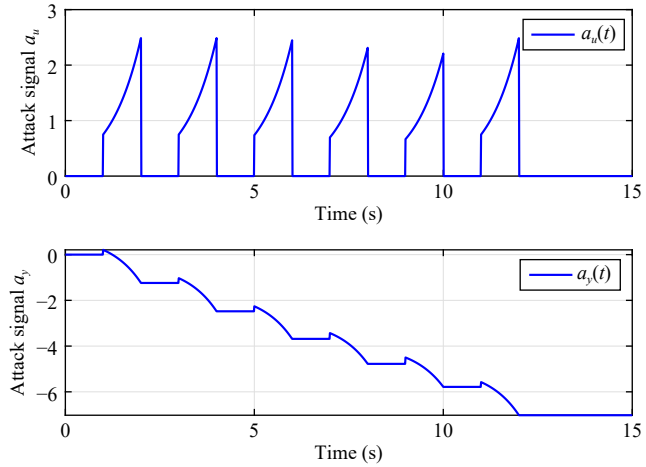


Fig. 4. Time responses of the attack signals $a_u(t)$ and $a_y(t)$.

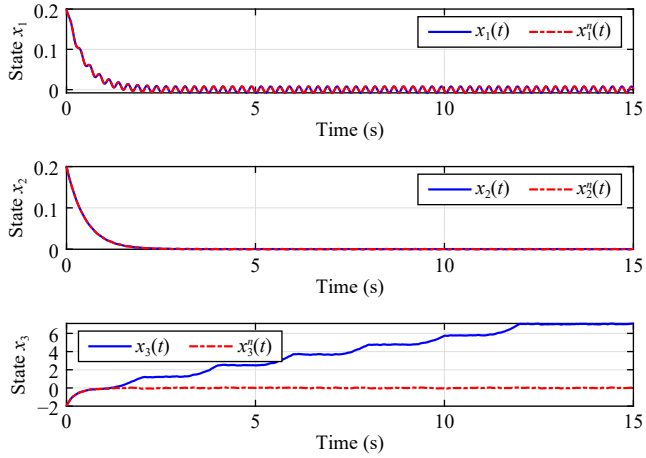


Fig. 5. Time responses of the state vector $x(t)$ in the attack case and $x^n(t)$ in the nominal case.

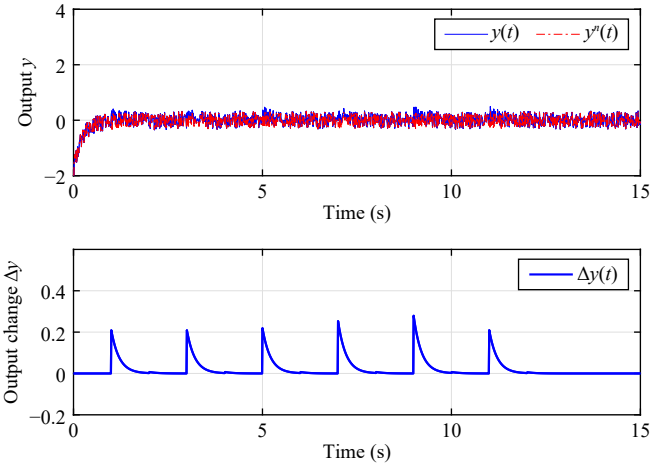


Fig. 6. Time responses of the sensor measurements $y(t)$ and $y^n(t)$ in the attack case and nominal case respectively, and the output change $\Delta y(t)$.

each of the attack resuming time instants {1 s, 3 s, 5 s, 7 s, 9 s, 11 s}, and such an increment converges to zero exponentially. Fig. 6 also shows that the increments caused by the attack are relatively small and hidden by the disturbances and noise, which is a result of the selected relatively "large" process disturbances and measurement noise used in the simulation. This particularly created simulation scenario is used for verifying that the designed backward-in-time detector is robust to the process disturbances and measurement noise, and also sensitive to the stealthy intermittent integrity attacks.

Fig. 7 illustrates the anomaly detection results using the equipped anomaly detector $\mathcal{D}$, in which the residual $r(t)$ remains far below the threshold J_{th} during the attack event. Hence, the attack is not detected by $\mathcal{D}$. This indicates that the change Δy due to the intermittent attack is sufficiently small to maintain the stealthiness of the attack with respect to the detector $\mathcal{D}$.

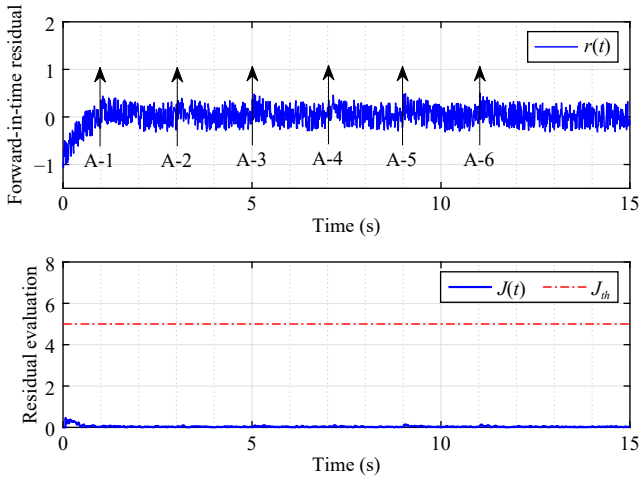


Fig. 7. Time responses of the forward-in-time residual $r(t)$, the evaluation function $J(t) = |r(t)|$ and the threshold J_{th} of the anomaly detector $\mathcal{D}$.

B. Attack Detection

Following Algorithm 1, the parameters for the SMOOTHER procedure are given as follows: $t_b = 2$, $T = 15$, $t_{s,0} = t_b = 2$, $t_{s,1} = 9$, $t_{s,2} = 16$, $\alpha = 0.009$ and $R = 0.1$. Moreover, Θ_k satisfying (49) is given by $\Theta_0 = 8.5I_3, \Theta_1 = 43.5I_3$. By solving the differential equations in (41), P_1 and Σ_1 are obtained as follows:

$$P_1 = \begin{bmatrix} 0.0162 & 0.0025 & 0.0484 \\ 0.0027 & 0.0008 & 0.0153 \\ 0.0544 & 0.0101 & 0.2307 \end{bmatrix}$$

$$\Sigma_1 = \begin{bmatrix} 0.0010 & 0 & 0 \\ 0 & 0.0182 & 0 \\ 0 & 0 & 29.232 \end{bmatrix}.$$

Furthermore, $P(t)$ and $\Sigma(t)$ for [2 s, 9 s) and [9 s, 16 s) can be obtained respectively by solving the differential equations in (41), and thus, the fixed-point smoother can be implemented. Then, given $y(t_b) = 0.1985$, $r(t_b) = -0.1332$ and the estimate $\hat{\phi}(t)$, the RESIDUAL procedure can be completed by following Algorithm 1, and the residual $\hat{r}(t_b|t)$ from (29) and its evaluation function $\hat{J}(t_b|t)$ from (54) can be obtained. We proceed with the THRESHOLD procedure based on Algorithm 1. The scalars $\bar{\omega}$ and $\bar{v}$ in Assumption 2 are given by $\bar{\omega} = 0.3$ and $\bar{v} = 0.2$. Moreover, $\beta_0 = 0.25$, $\lambda_0 = 0.64$, $\beta_1 = 0.1$ and $\lambda_1 = 0.63$, and $\delta_0 = 100$. The initial value of $\varepsilon_{\phi,k}$ is chosen as $\varepsilon_{\phi,0} = 5$. By following the THRESHOLD procedure in Algorithm 1, the threshold $\hat{J}_{th}(t)$ in (57) can be calculated.

The residual $\hat{r}(2|t)$, the evaluation function $\hat{J}(2|t)$ and the threshold $\hat{J}_{th}(t)$ are shown in Fig. 8. It is shown that at each of the attack resuming time instants {3 s, 5 s, 7 s, 9 s, 11 s}, the residual $\hat{r}(2|t)$ and its corresponding evaluation $\hat{J}(2|t)$ have a jump in magnitude. Thus, the accumulation property described in Lemma 1 is satisfied. Furthermore, note also that in Fig. 8, the jump that occurs at $t = 9$ s is much larger than the one at $t = 7$ s due to covariance matrix resetting, even though the increment Δy at $t = 9$ s is similar with the one at $t = 7$ s (see $\Delta z_4 = \Delta z_5$). Moreover, based on the DECISION PRINCIPLE specified in Algorithm 1, we can conclude from Fig. 8 that the injected intermittent stealthy integrity attack is successfully detected at the time $T_d \approx 9.5$ s when the evaluation function $\hat{J}(2|t)$ exceeds the threshold $\hat{J}_{th}(t)$.

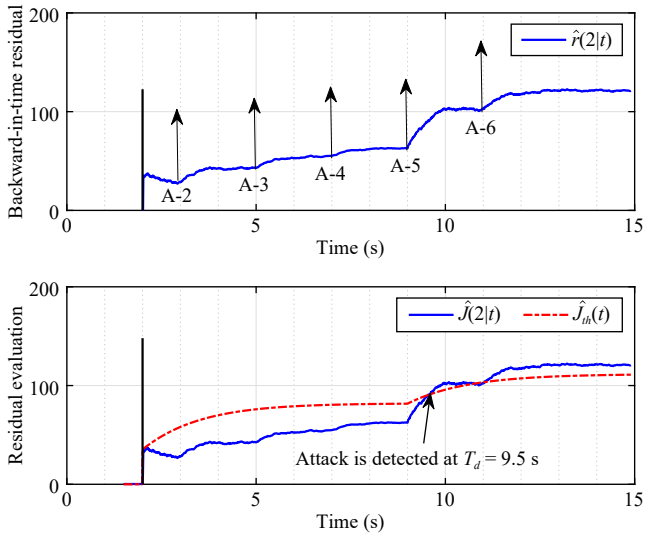


Fig. 8. Time responses of the backward-in-time residual $\hat{r}(2|t)$, the evaluation function $\hat{J}(2|t)$ and the threshold $\hat{J}_{th}(t)$.

VIII. Conclusion

In this paper, stealthy intermittent integrity attacks being stealthy with respect to typical anomaly detectors have been formulated. A backward-in-time detection residual that can accumulate at each attack activation time and is able to indicate the stealthy intermittent integrity attacks has been introduced. A fixed-point smoother has been designed as the backward-in-time estimator for estimating the unknown backward-in-time detection residual. A covariance matrix resetting technique has been applied in the design of the smoother to guarantee the required sensitivity to the attacks. The corresponding adaptive threshold generation scheme for detecting the stealthy intermittent integrity attacks has been designed, and the attack detectability has also been investigated rigorously. Some future research works are given as follows:

1) One of our studies work focuses on scheduling the attack

pausing and resuming time instants such that the attack can cause significant damage to the system, and at the same time, achieve the power energy saving aim. Game theory may provide a way to solve the trade-off between damaging effects and energy saving [38].

2) Another future research direction involves the modification of typical fixed-point smoothers to improve the sensitivity to stealthy integrity attacks. The forgetting factor for the covariance matrix used in [51] is a potential way to achieve this improvement.

3) Resilience control against intermittent integrity attacks is also one potential research direction. Some control issues, such as quantized sensor measurements well handled by the model reference control methodologies in [52], [53] and high order nonlinearities in [54], [55] will be investigated.

4) Stealthy intermittent integrity attack generation and detection issues for nonlinear systems and large-scale systems such as in [56] will be considered in our future research. Note that a way for generating stealthy intermittent attacks for nonlinear systems and large-scale systems is to use the geometric approach proposed in [14].

Appendix A
Proof of Theorem 1

Proof : Let Δx_1, Δx_2, Δy_1, Δy_2 and Δu represent the changes of x_1, x_2, y_1, y_2 and u respectively due to the attack $a(t)$, i.e., $\Delta x = x - x^n$, $\Delta x_2 = x_2 - x_2^n$, $\Delta y_1 = y_1 - y_1^n$, $\Delta y_2 = y_2 - y_2^n$ and $\Delta u = u - u^n$. Then, from (10) and (11), the incremental systems between $\mathcal{W}_j$ and $\mathcal{W}_j^n$ for $j = 1, 2$, denoted as $\Delta\mathcal{W}_j$, can be obtained as follows:

$$\Delta\mathcal{W}_1 : \begin{cases} \Delta\dot{x}_1(t) = A\Delta x_1(t) + B_a a(t) \\ \Delta y_1(t) = C\Delta x_1(t) + D_a a(t) \end{cases} \tag{66}$$

$$\Delta\mathcal{W}_2 : \begin{cases} \Delta\dot{x}_2(t) = A\Delta x_2(t) + B\Delta u(t) \\ \Delta y_2(t) = C\Delta x_2(t) \end{cases} \tag{67}$$

where $\Delta x_1(t_k) = -\Delta z_k$ and $\Delta x_2(t_k) = \Delta z_k$. Thus, the state change Δx and the output change Δy can be written as

$$\Delta x(t) = \Delta x_1(t) + \Delta z(t) \tag{68}$$

$$\Delta y(t) = \Delta y_1(t) + \Delta y_2(t). \tag{69}$$

Consider $\mathcal{W}_j$ for $j = 1, 2$ given in (11). Denote $\mathcal{W}_{j,k}^n$ as the "nominal" system of $\mathcal{W}_j$ in the absence of the attack signal $a_k(t)$ but in the presence of $a_1(t), \ldots, a_{k-1}(t)$, where $k \geq 1$. By letting $x_{1,k}^n$ and $y_{1,k}^n$ represent the state and the output of $\mathcal{W}_{1,k}^n$, and letting $x_{2,k}^n$, $y_{2,k}^n$ and u_k^n represent the state, the output and the control of $\mathcal{W}_{2,k}^n$, respectively, it then follows from (11) that $\mathcal{W}_{j,k}^n$ can be written as:

$$\mathcal{W}_{1,k}^n : \begin{cases} \dot{x}_{1,k}^n(t) = Ax_{1,k}^n(t) + \sum_{i=1}^{k-1} B_a a_k(t) \\ y_{1,k}^n(t) = Cx_{1,k}^n(t) + \sum_{i=1}^{k-1} D_a a_k(t) \end{cases} \tag{70}$$

$$\mathcal{W}_{2,k}^n : \begin{cases} \dot{x}_{2,k}^n(t) = Ax_{2,k}^n(t) + Bu_k^n(t) + D\omega(t) \\ y_{2,k}^n(t) = Cx_{2,k}^n(t) + v(t) \end{cases} \tag{71}$$

where the initial conditions can be chosen as $x_1(t_k) - x_{1,k}^n(t_k) = -\Delta z_k$ and $x_2(t_k) - x_{2,k}^n(t_k) = \Delta z_k$ such that $x_1(t_k) + x_2(t_k) = x_{1,k}^n(t_k) + x_{2,k}^n(t_k)$.

Let $\Delta x_{1,k}$ and $\Delta y_{1,k}$ represent the changes of $x_{1,k}^n$, and $y_{1,k}^n$, and $\Delta x_{2,k}$, $\Delta y_{2,k}$ and Δu_k represent the changes of $x_{2,k}^n$, $y_{2,k}^n$ and u_k^n respectively due to the attack $a_k(t)$, i.e., $\Delta x_{1,k} = x_1 - x_{1,k}^n$, $\Delta y_{1,k} = y_1 - y_{1,k}^n$, $\Delta x_{2,k} = x_2 - x_{2,k}^n$, $\Delta y_{2,k} = y_2 - y_{2,k}^n$ and $\Delta u_k = u - u_k^n$. Then, based on (70), (71) and (11), the incremental systems, denoted as $\Delta\mathcal{W}_{j,k}$, are obtained as

$$\Delta\mathcal{W}_{1,k} : \begin{cases} \Delta\dot{x}_{1,k}(t) = A\Delta x_{1,k}(t) + B_a a_k(t) \\ \Delta y_{1,k}(t) = C\Delta x_{1,k}(t) + D_a a_k(t) \end{cases} \tag{72}$$

$$\Delta\mathcal{W}_{2,k} : \begin{cases} \Delta\dot{x}_{2,k}(t) = A\Delta x_{2,k}(t) + B\Delta u_k(t) \\ \Delta y_{2,k}(t) = C\Delta x_{2,k}(t) \end{cases} \tag{73}$$

where $\Delta x_{1,k}(t_k) = -\Delta z_k$ and $\Delta x_{2,k}(t_k) = \Delta z_k$.

Therefore, based on the superposition principle for linear systems ($a(t)$ is the sum of $a_k(t)$ in (8c)), Δx_1, Δz and Δy can be written as

$$\Delta x_1(t) = \sum_{i=1}^{k} \Delta x_{1,i}(t),\ \Delta x_2(t) = \sum_{i=1}^{k} \Delta x_{2,i}(t) \tag{74}$$

$$\Delta y(t) = \sum_{i=1}^{k} (\Delta y_{1,i}(t) + \Delta y_{2,i}(t)),\ \forall\, t \in \Omega_k. \tag{75}$$

In the sequel, the responses of $\Delta\mathcal{W}_{j,k}$ in the attack activating time interval and the attack silence time interval are analyzed.

1) Activating Time Interval Ω_k^{ac}: In this time interval, $\Delta\mathcal{W}_{1,k}$ and $a_k(t)$ described by (8a) and (8b) respectively can be equivalently written in the coordinates $(\Delta\bar{x}_{1,k}, \zeta_k)$ with $\Delta\bar{x}_{1,k} = \Delta x_{1,k} - \zeta_k$ as follows:

$$\begin{aligned} \Delta\dot{\bar{x}}_{1,k}(t) &= A\Delta\bar{x}_{1,k}(t) \\ \dot{\zeta}_k(t) &= (A + B_a F_k)\zeta_k(t) \\ \Delta y_{1,k}(t) &= C\Delta\bar{x}_{1,k}(t) + (C + D_a F_k)\zeta_k(t) \end{aligned}$$

where $\Delta\bar{x}_{1,k}(t_k) = \Delta x_{1,k}(t_k) - \zeta_k(t_k) = 0$. Since $\Delta\bar{x}_{1,k}(t_k) = 0$, $\Delta\bar{x}_{1,k}(t) = 0$ for $t \in \Omega_k^{\mathrm{ac}}$, and hence, $\Delta x_{1,k}(t) = \zeta_k(t)$ for $t \in \Omega_k^{\mathrm{ac}}$. For $\Delta z_k \in \mathcal{V}_0$ and F_k satisfying (12a), we have

$$\zeta_k(t) \in \mathcal{V}_0,\ (C + D_a F_k)\zeta_k(t) = 0,\ \forall\, t \in \Omega_k^{\mathrm{ac}}.$$

Thus, we obtain

$$\Delta y_{1,k}(t) = 0,\ \forall\, t \in \Omega_k^{\mathrm{ac}}. \tag{76}$$

2) Silence Time Interval Ω_k^{si}: In this time interval, the initial condition of $\Delta\mathcal{W}_{1,k}$ is $\Delta x_{1,k}(t_k + \tau_k) = \Delta x_{1,k}((t_k + \tau_k)^-) \in \mathcal{V}_0$. Based on $a_k(t)$ in (8b) during $t \in \Omega_k^{\mathrm{si}}$, $\Delta x_{1,k}$ satisfies

$$\Delta\dot{x}_{1,k}(t) = A\Delta x_{1,k}(t),\ \Delta x_{1,k}(t_k + \tau_k) \in \mathcal{V}_0.$$

Also, $\Delta y_{1,k}$ at the time instant $t_k + \tau_k$ satisfies

$$\begin{aligned} \Delta y_{1,k}(t_k + \tau_k) &= C\Delta x_{1,k}(t) + a_{y,k}((t_k + \tau_k)^-) \\ &= \Delta y_{1,k}((t_k + \tau_k)^-) = 0. \end{aligned} \tag{77}$$

Thus, it follows from $\Delta x_{1,k}(t_k + \tau_k) \in \mathcal{V}_0$ and $\mathcal{V}_0 \subset \mathcal{H}$ with $\mathcal{H}$ being the unobservable subspace of the pair (CA, A) that $\Delta x_{1,k}(t_k + \tau_k) \in \mathcal{H}$, which indicates:

$$\Delta\dot{y}_{1,k}(t) = CA\Delta x_{1,k}(t) = 0,\ \forall\, t \in \Omega_k^{\mathrm{si}}. \tag{78}$$

Thus, by combining (77) and (78), we have

$$\Delta y_{1,k}(t) = 0,\ \forall\, t \in \Omega_k^{\mathrm{si}}. \tag{79}$$

Hence, from the result (76) in the activating time interval and the result (79) in the silence time interval, we can conclude

$$\Delta y_{1,k}(t) = 0,\ \forall\, t \in \Omega_k^{0}.$$

Therefore, it follows from (75) that the change Δy can be written as in (13) with $\Delta y_{2,k}(t)$ being generated by $\Delta\mathcal{W}_{2,k}$ in (79). In addition, since $\Delta u_k = K\Delta y_k = K\Delta y_{2,k} = KC\Delta z_k$, (79) can be written as (14). ■

Appendix B
Proof of Lemma 1

Proof:

1) By using the transition matrix Φ in (17), the solution of the system $\Delta\mathcal{W}$ in (15) with the initial condition $\Delta z(t_k) = \Delta z_k$ can be written as

$$\Delta z(t) = \Phi(t, t_k)\Delta z_k,\ \forall\, t \in \Omega_k.$$

Thus, from Definition 1, $\Delta z(t_b|t)$ can be written as

$$\Delta z(t_b|t) = \Phi(t_b, t)\Phi(t, t_k)\Delta z_k = \Phi(t_b, t_k)\Delta z_k,\ \forall\, t \in \Omega_k. \tag{80}$$

Since both $\Phi(t_b, t_k)$ and Δz_k are independent of time t, $\Delta z(t_b|t)$ is also independent of time t, then $\Delta z(t_b|t)$ is a constant vector with respect to time. Hence, (18) follows.

2) Since $\Delta z_k \neq 0$ and $\Phi(t_b, t_k) \neq 0$, then the result (19) follows directly from (80).

3) From (80), for the consecutive attack slots Ω_k and Ω_{k+1}, we have

$$\Delta z(t_b|t) = \Phi(t_b, t_k)\Delta z_k,\ \forall\, t \in \Omega_k$$
$$\Delta z(t_b|t) = \Phi(t_b, t_{k+1})\Delta z_{k+1},\ \forall\, t \in \Omega_{k+1}.$$

By using $\Phi(t_b, t_k) = \Phi(t_b, t_{k+1})\Phi(t_{k+1}, t_k)$, $\Delta z(t_b|t_k)$ can be equivalently written as

$$\Delta z(t_b|t) = \Phi(t_b, t_{k+1})\Phi(t_{k+1}, t_k)\Delta z_k,\ \forall\, t \in \Omega_k.$$

Then, we can derive

$$|\Delta z(t_b|t)|^2 \le \bar{\sigma}^2(\Phi(t_b, t_{k+1})|\Phi(t_{k+1}, t_k)\Delta z_k|^2,\ \forall\, t \in \Omega_k$$
$$|\Delta z(t_b|t)|^2 \ge \underline{\sigma}^2(\Phi(t_b, t_{k+1}))|\Delta z_{k+1}|^2,\ \forall\, t \in \Omega_{k+1}.$$

Hence, $|\Delta z(t_b|t)|_{t\in\Omega_{k+1}} \ge |\Delta z(t_b|t)|_{t\in\Omega_k}$ if (20) is satisfied. ■

Appendix C
Proof of Theorem 3

Proof:

1) Based on Lemma 1, in the absence of the attack, $\Delta z(t_b|t) = 0$ and from (21), $\Delta y(t_b|t) = 0$, and thus, $r(t_b|t) = r(t_b)$. Since $|r(t_b)| \le J_{th}$, result 1) follows.

2) By using the reverse triangle inequality and based on (21) and (22), a sufficient condition to guarantee $|r(t_b|t)| > J_{th}$ for all $t \in \bigcup_{i=k}^{N_a}\Omega_i$ can be obtained as

$$|C\Delta z(t_b|t)| > \frac{J_{th} + |r(t_b)|}{\underline{\sigma}(C)},\ \forall\, t \in \bigcup_{i=k}^{N_a}\Omega_i. \tag{81}$$

Note that based on result 1) in Lemma 1, $\Delta z(t_b|t)$ is a constant vector during an attack slot Ω_k. Thus, $|\Delta z(t_b|t)| = |\Delta z(t_b|t_k)|$ for $t \in \Omega_k$. Note also that under Assumption 1 and based on result 3) in Lemma 1, $|\Delta z(t_b|t_i)| \ge |\Delta z(t_b|t_k)|$ for any $i \ge k$ since (20) is considered to hold. Thus, a sufficient condition to guarantee (81) is obtained as follows:

$$|\Delta z(t_b|t_k)| > \frac{J_{th} + |r(t_b)|}{\underline{\sigma}(C)}. \tag{82}$$

Based on Definition 1, we can write $\Delta z(t_b|t_k) = \Phi^{-1}(t_k, t_b)\Delta z_k$. Thus, a sufficient condition to guarantee (82) can be obtained as

$$|\Phi(t_k, t_b)| < \frac{\underline{\sigma}(C)|\Delta z_k|}{J_{th} + |r(t_b)|}. \tag{83}$$

Note that according to [40], for the Hurwitz matrix $A + BKC$, there exist $k_0 > 0$ and $\lambda_0 > 0$ such that

$$|\Phi(t_k, t_b)| \le k_0 e^{-\lambda_0(t_k - t_b)}.$$

Thus, the time t_k satisfying (25) can guarantee the sufficient condition (83). Hence, result 2) follows. ■

References

[1] A. A. Cardenas, S. Amin, and S. Sastry, "Secure control: Towards survivable cyber-physical systems," in *Proc. 28th Int. Conf. Distributed Computing Systems Workshops*, Beijing, China, 2008, pp. 495–500.

[2] S. M. Dibaji, M. Pirani, D. B. Flamholz, A. M. Annaswamy, K. H. Johansson, and A. Chakrabortty, "A systems and control perspective of CPS security," *Annu. Rev. Control*, vol. 47, pp. 394–411, Jan. 2019.

[3] V. L. Do, L. Fillatre, I. Nikiforov, and P. Willett, "Security of SCADA systems against cyber-physical attacks," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 32, no. 5, pp. 28–45, May 2017.

[4] A. Hobbs, "The colonial pipeline hack: Exposing vulnerabilities in U.S. cybersecurity," 2021. [Online]. Available: https://sk.sagepub.com/cases/colonial-pipeline-hack-exposing-vulnerabilities-us-cybersecurity.

[5] W. L. Duo, M. C. Zhou, and A. Abusorrah, "A survey of cyber attacks on cyber physical systems: Recent advances and challenges," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 5, pp. 784–800, May 2022.

[6] Y. L. Mo and B. Sinopoli, "Secure control against replay attacks," in *Proc. 47th Annu. Allerton Conf. Communication, Control, and Computing*, Monticello, USA, 2009, pp. 911–918.

[7] R. S. Smith, "Covert misappropriation of networked control systems: Presenting a feedback structure," *IEEE Control Syst. Mag.*, vol. 35, no. 1, pp. 82–92, Feb. 2015.

[8] A. Barboni, H. Rezaee, F. Boem, and T. Parisini, "Detection of covert cyber-attacks in interconnected systems: A distributed model-based approach," *IEEE Trans. Autom. Control*, vol. 65, no. 9, pp. 3728–3741, Sept. 2020.

[9] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "A secure control framework for resource-limited adversaries," *Automatica*, vol. 51, pp. 135–148, Jan. 2015.

[10] Q. R. Zhang, K. Liu, D. Y. Han, G. Z. Su, and Y. Q. Xia, "Design of stealthy deception attacks with partial system knowledge," *IEEE Trans. Autom. Control*, vol. 68, no. 2, pp. 1069–1076, Feb. 2023.

[11] A. Teixeira, K. C. Sou, H. Sandberg, and K. H. Johansson, "Secure control systems: A quantitative risk management approach," *IEEE Control Syst. Mag.*, vol. 35, no. 1, pp. 24–45, Feb. 2015.

[12] H. S. Sánchez, D. Rotondo, T. Escobet, V. Puig, and J. Quevedo, "Bibliographical review on cyber attacks from a control oriented perspective," *Annu. Rev. Control*, vol. 48, pp. 103–128, Sept. 2019.

[13] T. Y. Zhang and D. Ye, "False data injection attacks with complete stealthiness in cyber-physical systems: A self-generated approach," *Automatica*, vol. 120, p. 109117, Oct. 2020.

[14] K. K. Zhang, C. Keliris, T. Parisini, and M. M. Polycarpou, "Stealthy integrity attacks for a class of nonlinear cyber-physical systems," *IEEE*

Trans. Autom. Control, vol. 67, no. 12, pp. 6723–6730, Dec. 2022.

[15] A. Y. Lu and G. H. Yang, "Input-to-state stabilizing control for cyber-physical systems with multiple transmission channels under denial of service," *IEEE Trans. Autom. Control*, vol. 63, no. 6, pp. 1813–1820, 2018.

[16] H. Zhang, P. Cheng, L. Shi, and J. M. Chen, "Optimal denial-of-service attack scheduling with energy constraint," *IEEE Trans. Autom. Control*, vol. 60, no. 11, pp. 3023–3028, Nov. 2015.

[17] H. Zhang, Y. F. Qi, J. F. Wu, L. K. Fu, and L. D. He, "DoS attack energy management against remote state estimation," *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 1, pp. 383–394, Mar. 2018.

[18] S. Amin, A. A. Cárdenas, and S. S. Sastry, "Safe and secure networked control systems under denial-of-service attacks," in *Proc. 12th Int. Workshop on Hybrid Systems: Computation and Control*, San Francisco, USA, 2009, pp. 31–45.

[19] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "Revealing stealthy attacks in control systems," in *Proc. 50th Annu. Allerton Conf. Communication, Control, and Computing*, Monticello, USA, 2012, pp. 1806–1813.

[20] F. Pasqualetti, F. Dörfler, and F. Bullo, " Attack detection and identification in cyber-physical systems," *IEEE Trans. Autom. Control*, vol. 58, no. 11, pp. 2715–2729, Nov. 2013.

[21] R. S. Smith, "A decoupled feedback structure for covertly appropriating networked control systems," *IFAC Proc. Vol.*, vol. 44, no. 1, pp. 90–95, Jan. 2011.

[22] Y. B. Mao, H. Jafarnejadsani, P. Zhao, E. Akyol, and N. Hovakimyan, "Novel stealthy attack and defense strategies for networked control systems," *IEEE Trans. Autom. Control*, vol. 65, no. 9, pp. 3847–3862, Sept. 2020.

[23] Y. L. Mo, R. Chabukswar, and B. Sinopoli, "Detecting integrity attacks on SCADA systems," *IEEE Trans. Control Syst. Technol.*, vol. 22, no. 4, pp. 1396–1407, Jul. 2014.

[24] R. M. G. Ferrari and A. M. H. Teixeira, " A switching multiplicative watermarking scheme for detection of stealthy cyber-attacks," *IEEE Trans. Autom. Control*, vol. 66, no. 6, pp. 2558–2573, Jun. 2021.

[25] A. Hoehn and P. Zhang, "Detection of covert attacks and zero dynamics attacks in cyber-physical systems," in *Proc. American Control Conf.*, Boston, USA, 2016, pp. 302–307.

[26] S. Weerakkody and B. Sinopoli, "Detecting integrity attacks on control systems using a moving target approach," in *Proc. 54th IEEE Conf. Decision and Control*, Osaka, Japan, 2015, pp. 5820–5826.

[27] P. Griffioen, S. Weerakkody, and B. Sinopoli, "A moving target defense for securing cyber-physical systems," *IEEE Trans. Autom. Control*, vol. 66, no. 5, pp. 2016–2031, May 2021.

[28] M. M. Polycarpou and A. J. Helmicki, "Automated fault detection and accommodation: A learning systems approach," *IEEE Trans. Syst. Man Cybern.*, vol. 25, no. 11, pp. 1447–1458, Nov. 1995.

[29] S. X. Ding, *Model-Based Fault Diagnosis Techniques: Design Schemes, Algorithms, and Tools*. 2nd ed. London, UK: Springer, 2013.

[30] M. Blanke, M. Kinnaert, J. Lunze, and M. Staroswiecki, *Diagnosis and Fault-Tolerant Control*. 2nd ed. Berlin, Germany: Springer, 2006.

[31] Y. K. Wu, B. Jiang, and N. Y. Lu, "A descriptor system approach for estimation of incipient faults with application to high-speed railway traction devices," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 49, no. 10, pp. 2108–2118, Oct. 2019.

[32] K. K. Zhang, B. Jiang, X. G. Yan, and Z. H. Mao, " Incipient fault detection for traction motors of high-speed railways using an interval sliding mode observer," *IEEE Trans. Intell. Transport. Syst.*, vol. 20, no. 7, pp. 2703–2714, Jul. 2019.

[33] C. Keliris, M. M. Polycarpou, and T. Parisini, "An integrated learning and filtering approach for fault diagnosis of a class of nonlinear dynamical systems," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 28, no. 4, pp. 988–1004, Apr. 2017.

[34] M. Taheri, K. Khorasani, I. Shames, and N. Meskin, Cyber Attack and machine induced fault detection and isolation methodologies for cyber-physical systems, 2020. [Online]. Available: https://arxiv.org/abs/2009.06196.

[35] K. K. Zhang, M. M. Polycarpou, and T. Parisini, "Enhanced anomaly detector for nonlinear cyber-physical systems against stealthy integrity attacks," *IFAC-PapersOnLine* , vol. 53, no. 2, pp. 13682–13687, Jan. 2020.

[36] K. K. Zhang, C. Keliris, M. M. Polycarpou, and T. Parisini, "Detecting stealthy integrity attacks in a class of nonlinear cyber-physical systems: A backward-in-time approach," *Automatica* , vol. 141, p. 110262, Jul. 2022.

[37] E. Kontouras, A. Tzes, and L. Dritsas, "Hybrid detection of intermittent cyber-attacks in networked power systems," *Energies*, vol. 12, no. 24, p. 4625, Dec. 2019.

[38] S. Gao, H. Zhang, Z. P. Wang, C. Huang, and H. C. Yan, "Optimal injection attack strategy for cyber-physical systems under resource constraint: A game approach," *IEEE Trans. Control Netw. Syst.*, to be published.

[39] J. Chen and R. J. Patton, *Robust Model-Based Fault Diagnosis for Dynamic Systems*. Springer, 2012.

[40] X. D. Zhang, M. M. Polycarpou, and T. Parisini, "Fault diagnosis of a class of nonlinear uncertain systems with Lipschitz nonlinearities using adaptive estimation," *Automatica*, vol. 46, no. 2, pp. 290–299, 2010.

[41] K. K. Zhang, B. Jiang, X. G. Yan, and J. Shen, "Interval sliding mode observer based incipient sensor fault detection with application to a traction device in China railway high-speed," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2585–2597, 2019.

[42] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*. Englewood Cliffs: Prentice-Hall, 1993.

[43] B. D. Anderson and J. B. Moore, *Optimal Filtering*. North Chelmsford, USA: Courier Corporation, 2012.

[44] M. Green and D. J. N. Limebeer, *Linear Robust Control*. New York, USA: Dover Publications, 2012.

[45] D. Simon, *Optimal State Estimation: Kalman, H_∞ , and Nonlinear Approaches*. Hoboken, USA: John Wiley & Sons, 2006.

[46] G. A. Einicke, *Smoothing, Filtering and Prediction: Estimating the Past, Present and Future*. Rijeka: IntechOpen, 2012.

[47] X. B. Li and K. M. Zhou, " A time domain approach to robust fault detection of linear time-varying systems," *Automatica* , vol. 45, no. 1, pp. 94–102, Jan. 2009.

[48] R. N. Banavar and J. L. Speyer, "A linear-quadratic game approach to estimation and smoothing," in *Proc. American Control Conf.*, Boston, USA, 1991, pp. 2818–2822.

[49] H. Abou-Kandil, G. Freiling, V. Ionescu, and G. Jank, *Matrix Riccati Equations in Control and Systems Theory*. Birkhäuser Verlag, Basel, 2012.

[50] G. Basile and G. Marro, *Controlled and Conditioned Invariants in Linear System Theory*. Englewood Cliffs: Prentice Hall, 1992.

[51] Q. J. Xia, M. Rao, Y. Q. Ying, and X. M. Shen, " Adaptive fading Kalman filter with an application," *Automatica*, vol. 30, no. 8, pp. 1333–1338, Aug. 1994.

[52] Y. J. Zhang, J. F. Zhang, X. K. Liu, and Z. Liu, "Quantized-output feedback model reference control of discrete-time linear systems," *Automatica*, vol. 137, p. 110027, Mar. 2022.

[53] J. Guo, Y. J. Zhang, J. F. Zhang, and X. K. Liu, " Finite quantized-output feedback tracking control of possibly non-minimum phase linear systems," *IEEE Control Syst. Lett.*, vol. 6, pp. 2407–2412, Mar. 2022.

[54] M. L. Lv, W. W. Yu, J. D. Cao, and S. Baldi, " A separation-based methodology to consensus tracking of switched high-order nonlinear multiagent systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 5467–5479, Oct. 2022.

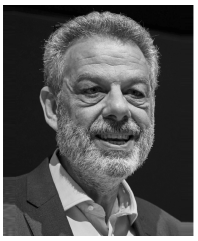
[55] M. L. Lv, B. De Schutter, C. Shi, and S. Baldi, "Logic-based distributed switching control for agents in power-chained form with multiple unknown control directions," *Automatica* , vol. 137, p. 110143, Mar. 2022.

[56] Y. Liu, D. Y. Yao, L. J. Wang, and S. J. Lu, " Distributed adaptive fixed-time robust platoon control for fully heterogeneous vehicles," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 53, no. 1, pp. 264–274, Jan. 2023.

Kangkang Zhang received the B.Eng., M.Sc. and Ph.D. degrees in control science and engineering from Henan University of Technology in 2012, Northeastern University in 2014, Nanjing University of Aeronautics and Astronautics (NUAA) in 2018, respectively. During 2017−2018, he was a visiting Ph.D. candidate at University of Kent, UK. From 2019 to 2022, he was a Research Associate at the KIOS Research and Innovation Center, University of Cyprus, Cyprus. He currently works as a Marie SkBodowska-Curie Individual Fellow at the Control and Power Group, Imperial College London, UK. His research interests include fault diagnosis, security of cyber-physical systems and fault-tolerant control. Dr. Zhang is a reviewer for various conferences and journals. His thesis received the CAA Excellent Doctoral Dissertation Award from Chinese Association of Automation.

Christodoulos Keliris received the Diploma (Hons.) degree in electrical and computer engineering from the Aristotle University of Thessaloniki, Greece, in 2007, the M.Sc. (Hons.) degree in finance from Imperial College London, UK in 2008, and the Ph.D. degree in electrical engineering from the University of Cyprus, Cyprus, in 2015. He was a Researcher in various European research and operational programmes. His current research interests include fault diagnosis for nonlinear systems, nonlinear control theory, adaptive learning, intelligent systems, and security of cyber-physical systems.

Dr. Keliris is a Reviewer for various conferences and journals. His dissertation Pricing Barrier Options received the Best M.Sc. Finance Dissertation Prize.



Thomas Parisini (Fellow, IEEE) received the Ph.D. degree in electronic engineering and computer science in 1993 from the University of Genoa, Italy. He was with Politecnico di Milano and since 2010. He holds the Chair of Industrial Control and is Director of Research at Imperial College London, UK. He is a Deputy Director of the KIOS Research and Innovation Centre of Excellence, University of Cyprus, Cyprus. Since 2001 he is also Danieli Endowed Chair of Automation Engineering with University of Trieste, Italy. In 2009−2012 he was Deputy Rector of University of Trieste, Italy. In 2018 he received an Honorary Doctorate from University of Aalborg, Denmark. In 2021, he served as Deputy Chair of the Employment & Education Task Force of the B20, Italy. He authored or coauthored more than 350 research papers in archival journals, book chapters, and international conference proceedings. He is a co-recipient of the IFAC Best Application Paper Prize of the Journal of Process Control, Elsevier, for the three year period 2011−2013 and of the 2004 Outstanding Paper Award of the IEEE Trans. on Neural Networks. He is also a recipient of the 2007 IEEE Distinguished Member Award. In 2016, he was awarded as Principal Investigator at Imperial of the H2020 European Union flagship Teaming Project KIOS Research and Innovation Centre of Excellence led by University of Cyprus, Cyprus. Thomas Parisini serves as 2021−2022 President of the IEEE Control Systems Society and has served as Vice-President for Publications Activities. During 2009−2016 he was the Editor-in-Chief of the *IEEE Transactions on Control Systems Technology*. Since 2017, he is Editor for *Control Applications of Automatica* and since 2018 he is the Editor in Chief of the *European Journal of Control*. Among other activities, he was the Program Chair of the 2008 IEEE Conference on Decision and Control and General Co-Chair of the 2013 IEEE Conference on Decision and Control. Prof. Parisini is a Fellow of the IEEE and of the IFAC.

Bin Jiang (Fellow, IEEE) received the Ph.D. degree in automatic control from Northeastern University in 1995. He had ever been a Post-Doctoral Fellow, a Research Fellow, an Invited Professor, and a Visiting Professor in Singapore, France, USA, and Canada, respectively.

He is currently a Chair Professor of the Cheung Kong Scholar Program with the Ministry of Education and the Vice President of the Nanjing University of Aeronautics and Astronautics. He has authored eight books and over 100 referred international journal articles. His current research interests include intelligent fault diagnosis and fault tolerant control and their applications to helicopters, satellites, and high-speed trains. He is a fellow of the Chinese Association of Automation (CAA). He was a recipient of the Second-Class Prize of National Natural Science Award of China. He currently serves as an Editor for *International Journal of Control, Automation and Systems*, an Associate Editor or an Editorial Board Member for a number of journals, such as the *IEEE Transactions on Cybernetics*, and *IEEE Transactions on Neural Networks and Learning Systems*. He is also a Chair of Control Systems Chapter in IEEE Nanjing Section, and a member of the IFAC Technical Committee on Fault Detection, Supervision, and Safety of Technical Processes.

Marios M. Polycarpou (Fellow, IEEE) is a Professor of electrical and computer engineering and the Director of the KIOS Research and Innovation Center of Excellence at the University of Cyprus, Cyprus. He is also a Member of the Cyprus Academy of Sciences, Letters, and Arts, and an Honorary Professor of Imperial College London. He received the B.A degree in computer science and the B.Sc. in electrical engineering, both from Rice University, USA in 1987, and the M.S. and Ph.D. degrees in electrical engineering from the University of Southern California, USA, in 1989 and 1992, respectively. His teaching and research interests include intelligent systems and networks, adaptive and learning control systems, fault diagnosis, machine learning, and critical infrastructure systems. Dr. Polycarpou has published more than 400 articles in refereed journals, edited books and refereed conference proceedings, and co-authored 7 books. He is also the holder of 6 patents.

Prof. Polycarpou received the 2016 IEEE Neural Networks Pioneer Award. He is a Fellow of IEEE and IFAC and the recipient of the 2014 Best Paper Award for the journal Building and Environment (Elsevier). He served as the President of the IEEE Computational Intelligence Society (2012–2013), as the President of the European Control Association (2017–2019), and as the Editor-in-Chief of the *IEEE Transactions on Neural Networks and Learning Systems* (2004–2010). Prof. Polycarpou serves on the Editorial Boards of the Proceedings of the IEEE, the Annual Reviews in Control, and the Foundations and Trends in Systems and Control. His research work has been funded by several agencies and industry in Europe and the United States, including the prestigious European Research Council (ERC) Advanced Grant, the ERC Synergy Grant and the EU Teaming project. Prof. Polycarpou is the recipient of the 2023 IEEE Frank Rosenblatt Technical Field Award.