Detecting Stealthy Integrity Attacks in a Class of Nonlinear Cyber-Physical Systems: A Backward-in-Time Approach ⋆

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

[a] *KIOS Research and Innovation Center of Excellence*

[b] *Dept. of Electrical and Computer Engineering, University of Cyprus, Nicosia, 1678, Cyprus*

[c] *Dept. of Electrical and Electronic Engineering, Imperial College London, London, SW7 2AZ, U.K.*

[d] *Dept. of Engineering and Architecture, University of Trieste, Trieste, 34127, Italy*

Abstract

This paper proposes a stealthy integrity attack detection methodology for a class of nonlinear cyber-physical systems subject to disturbances. An equivalent increment of the system at a time prior to the attack occurrence time is introduced, which is theoretically proved to be effective to detect stealthy integrity attacks. A backward-in-time estimator is developed via the fixed-point smoother design tool to exploit this equivalent increment and allow the detection of the attack. More specifically, an asymptotically stable incremental system is introduced to characterize stealthy integrity attacks, and its backward-in-time solution at a fixed time prior to the attack occurrence formulates the equivalent increment. When running reversely in time, the divergence property of such an asymptotically stable incremental system enables the equivalent increment to detect stealthy integrity attacks. A fixed-point smoother is introduced to estimate the unknown equivalent increment for a class of Lipschitz nonlinear physical plants, such that the estimation error satisfies the $\mathcal{H}_\infty$ performance objective. Based on the equivalent increment and its estimation provided by the smoother, suitable residual and threshold signals are designed that allow the detection of the attack, and a detectability analysis is conducted to rigorously characterize the class of detectable attacks. Finally, a case study is presented to illustrate the effectiveness of the developed backward-in-time attack detection methodology.

Key words: Stealthy attacks, nonlinear cyber-physical systems, attack detection, fixed-point smoother.

1 Introduction

Cyber-physical systems (CPS) have attracted significant attention as a result of their wide applications, including electric power transmission and distribution systems, water and gas distribution systems and transportation systems. CPS are complex systems, consolidating computing and communication capabilities with monitoring and control of physical entities (Cardenas, Amin, and Sastry (2008)). Industrial control systems (ICS) operated through Supervisory Control and Data Acquisition (SCADA) systems are typical examples. Unfortunately, vulnerabilities to malicious cyber threats may deteriorate greatly the smooth operation of these systems precisely due to the integration of cyber and physical entities (Pasqualetti, Dorfler, and Bullo (2015)). Several cyber attack events in CPS have taken place in recent years, such as the Stuxnet worm attack on Iranian nuclear facilities, the GPS spoofing attack on an American unmanned aerial vehicle RQ-170 operated by the United States Air Force and the Ukraine attack on Ukrainian power distribution networks (for more details, the reader can refer to Dibaji et al. (2019) and the references therein). Therefore, motivated by increasing security and safety demands, advanced malicious cyber attack detection technologies are urgently required.

Cyber attacks are usually based on rational adversary models for empowering intelligence and intent. In the past decade, several survey papers such as Teixeira, Shames, Sandberg, and Johansson (2015a);

⋆ This paper was not presented at any IFAC meeting. Corresponding author Kangkang Zhang.

Email addresses: zhang.kangkang@ucy.ac.cy (Kangkang Zhang), keliris.chris@gmail.com (Christodoulos Keliris), mpolycar@ucy.ac.cy (Marios M. Polycarpou), t.parisini@gmail.com (Thomas Parisini).

Teixeira, Sou, Sandberg, and Johansson (2015b); Sánchez, Rotondo, Escobet, Puig, and Quevedo (2019); Dibaji, Pirani, Flamholz, Annaswamy, Johansson, and Chakrabortty (2019), have provided overviews of the research on cyber attacks from a control perspective. *Integrity attacks*, including replay attacks (Mo and Sinopoli (2009)), covert attacks (Smith (2015); Barboni, Rezaee, Boem, and Parisini (2020)), zero-dynamics attacks (Teixeira, Shames, Sandberg, and Johansson (2015a)), false-data injection attacks (Zhang and Ye (2020); An and Yang (2017)), are based on compromising sensor and actuator data transmission networks by using malicious signals. Such integrity attacks are the most researched type of cyber attacks in the systems and control literature.

In terms of stealthiness, traditional anomaly detectors such as fault diagnosis schemes in Ding (2013); Chen and Patton (1999); Blanke, Kinnaert, Lunze, Staroswiecki, and Schröder (2006) may not be able to detect integrity attacks because they are by design stealthy to such detectors. Anomaly detectors have been well-established, aiming to decide whether the behavior of the monitored system is healthy or faulty, and then, to identify the source of the anomalous behavior. However, it is difficult to directly extend current research results in anomaly detection to the case of malicious attack detection. One reason for this is the inherent limitation of anomaly detectors, such as the usually large amplitude requirements for the fault or attack signal so that sufficient discrepancy is created that is able to be detected. If a malicious attack is intelligently designed to generate residuals with sufficiently small amplitude, then the attack can bypass anomaly detectors without being detected. Another reason is that integrity attacks affect the system behavior in a specially designed way. Integrity attacks intelligently compromise the smooth operation of the CPS using particularly designed attack signals, such that the outputs remain unchanged in the presence of the attack, which indicates that anomaly detectors based on system outputs may not be able to detect these attacks. Therefore, detecting stealthy integrity attacks presents a key challenge in cyber-physical system security.

In the past decade, some model-based attack detection methods have been proposed. Optimal attack detectors in terms of probability have been proposed in Mo et al. (2013); Ye and Zhang (2019). Adding watermarks to the control inputs and then detecting them is the main idea for detecting replay attacks in Mo and Sinopoli (2009) and Romagnoli, Weerakkody, and Sinopoli (2019). However, additive watermarks may cause control performance degradation. In order to deal with this drawback, the authors in Ferrari and Teixeira (2017) propose a sensor multiplicative watermark for detecting and isolating replay attacks, and such a result is extended in Ferrari and Teixeira (2021) to detect stealthy cyber attacks. Another detection strategy based on modifying the dynamics of the CPS loop is presented in Teixeira, Shames, Sandberg, and Johansson (2012); Hoehn and Zhang (2016); Griffioen, Weerakkody, and Sinopoli (2021). Because covert attacks and zero-dynamics attacks rely highly on accurate knowledge of the dynamics of the physical plants, such modifying closed-loop dynamics approach provides an effective way to reveal these attacks. In the case of linear time-invariant systems, Teixeira, Shames, Sandberg, and Johansson (2012) provide a method to detect zero-dynamics attacks by modifying the physical system structure. A moving target approach is proposed in Weerakkody and Sinopoli (2015) using a linear time-varying system as the moving target, and a nonlinear moving target is developed in Griffioen, Weerakkody, and Sinopoli (2021). In addition, Hoehn and Zhang (2016) design a constant modulation and a periodic modulation to change the paths of the control data transmission. However, modifying the dynamics may be difficult to implement in practice and may degrade the performance of the CPS. Another way to detect integrity attacks is proposed in Chen, Kar, and Moura (2016) by using initial state information. However, the requirement for the available initial conditions of the states of the system limits its practical application. System nonlinearities are usually overseen by the aforementioned literature. Moreover, control performance of the original closed-loop CPS are usually negatively affected by the detection measures such as watermarks and moving targets.

In this paper, a backward-in-time detection methodology is proposed for a class of CPS with nonlinear physical plant by using only an analytical redundancy approach, which does not affect the control performance of the original closed-loop CPS. The main idea behind the proposed method is to exploit the estimated system changes due to attacks at a fixed time prior to the attack occurrence time. The design tool is an optimal fixed-point smoother which in this paper is referred to as the backward-in-time estimator. In particular, by investigating the asymptotic output-zeroing strategy for generating stealthy integrity attacks, an asymptotically stable incremental system due to such attacks is derived for the class of nonlinear CPS considered in this paper. An equivalent increment at a time prior to the attack occurrence time is also formulated as the solution of such an incremental system. Based on the divergence property of such solution running reversely in time, the feasibility to detect stealthy integrity attacks is rigorously investigated. However, such an equivalent increment is unknown and should be estimated by using current time measurements. Taking advantage of the backward-in-time estimation characteristics of fixed-pointed smoothers, a fixed-point smoother is introduced to estimate such an unknown equivalent increment. This is done by considering that the physical plant belongs to a class of Lipschitz nonlinear systems. In order to possess finite time horizon $\mathcal{H}_\infty$ performance, the design parameters of the smoother are formulated as the solution of a differential Riccati equation. Suitable residual and

threshold signals are then designed based on the aforementioned equivalent increment and its estimation provided by the smoother, which allows the detection of the stealthy integrity attacks. Finally, a detectability analysis is rigorously conducted, which characterizes quantitatively the class of detectable attacks.

The rest of the paper is organized as follows. In Section 2, the problem formulation is given. In Section 3, the equivalent change is defined, and its feasibility to detect stealthy integrity attacks is presented. In Section 4, the details of the backward-in-time detection methodology along with the detectability analysis are described and in Section 5, a case study is presented. Finally, some conclusions are drawn in Section 6.

Notation: Consider a vector signal $x(t) : \mathbb{R}_{\geq 0} \to \mathbb{R}^n$. Then, $x(t) \equiv 0$ for $t \in [t_1, t_2] \subset \mathbb{R}_{\geq 0}$ means that $x(t) = 0$ identically for all $t \in [t_1, t_2]$; $x(t) \not\equiv 0$ for $t \in [t_1, t_2] \subset \mathbb{R}_{\geq 0}$ means that $x(t) \neq 0$ for at least one time instant $t \in [t_1, t_2]$. The 2-norm on a time interval $[t_1, t_2]$, and the root-mean-square (RMS) of the signal vector $x(t)$ are defined as $\|x(t)\|_{[t_1,t_2]} = (\int_{t_1}^{t_2} x^T(\tau)x(\tau)d\tau)^{\frac{1}{2}}$ and $\|x(t)\|_{\text{RMS}} = (\frac{1}{T_w}\int_{t-T_w}^{t} x^T(\tau)x(\tau)d\tau)^{\frac{1}{2}}$, where $T_w > 0$ is the length of the time window. For $t \in [t_1, t_2]$, we have $\|x(t)\|_{\text{RMS}} \leq \frac{1}{\sqrt{T_w}}\|x(t)\|_{[t_1,t_2]}$. For a constant vector $x \in \mathbb{R}^n$, $\|x\|$ represents the Euclidean norm and is defined as $\|x\| = (x^T x)^{\frac{1}{2}}$. Thus, for the constant vector x, we have $\|x\| = \|x\|_{\text{RMS}}$. A continuous function $\alpha : [0, a) \to [0, \infty)$ is said to belong to class $\mathcal{K}$ function if it is strictly increasing and $\alpha(0) = 0$. It is said to belong to class $\mathcal{K}_\infty$ if $a = \infty$ and $\alpha(r) \to \infty$ as $r \to \infty$. A continuous function $\beta : [0, a) \times [0, +\infty) \to [0, \infty)$ is said to belong to class $\mathcal{KL}$ function if, for each fixed s, the mapping $\beta(r, s)$ belongs to class $\mathcal{K}$ function with respect to r and for each fixed r, the mapping $\beta(r, s)$ is decreasing with respect to s and $\beta(r, s) \to 0$ as $s \to \infty$.

2 Problem Formulation

A general CPS subject to integrity type of cyber attacks is shown in Fig. 1. It consists of a physical plant $\mathcal{P}$, a feedback controller $\mathcal{C}$, an anomaly detector $\mathcal{D}$, an actuator communication network $\mathcal{N}_a$ and a sensor communication network $\mathcal{N}_s$. During an integrity cyber attack event, the attack generation block attempts to compromise the communication networks $\mathcal{N}_a$ and $\mathcal{N}_s$ by injecting false data $a_u(t)$ and $a_y(t)$ respectively. In such an attack scenario, the outputs of $\mathcal{N}_a$ and $\mathcal{N}_s$ are respectively described by

$$\mathcal{N}_a : \ \tilde{u}(t) = u(t) + \Gamma_u a_u(t), \tag{1a}$$
$$\mathcal{N}_s : \ \tilde{y}(t) = y(t) + \Gamma_y a_y(t), \tag{1b}$$

where $\tilde{u} \in \mathbb{R}^{n_u}$ is the control data received by the plant $\mathcal{P}$, $u \in \mathbb{R}^{n_u}$ is the control data computed by

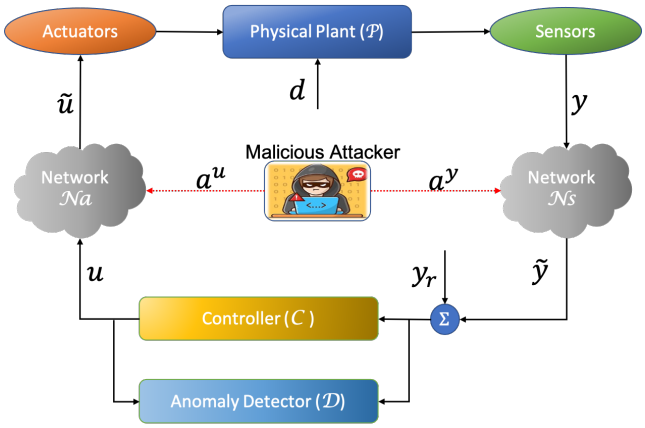

Fig. 1. General architecture of CPS under potential integrity cyber attacks.

the controller, $\tilde{y} \in \mathbb{R}^{n_y}$ is the sensor measurements received by the controller and the anomaly detector, and $y \in \mathbb{R}^{n_y}$ is the sensor measurements of the plant outputs transmitted by $\mathcal{N}_s$. Let $K_u \subseteq \{1, \cdots, n_u\}$ and $K_y \subseteq \{1, \cdots, n_y\}$ represent the disruption resources, i.e., the set of actuator and sensor communication channels that can be affected by the adversary. The distribution matrices $\Gamma_u \in \mathbb{B}^{n_u \times |K_u|}$ and $\Gamma_y \in \mathbb{B}^{n_y \times |K_y|}$ ($\mathbb{B} \triangleq \{0, 1\}$) are the binary incidence matrices mapping the attack signal to the respective channels. The attack signals are $a_u(t) = [a_{u,1}(t), \cdots, a_{u,|K_u|}(t)]^T \in \mathbb{R}^{|K_u|}$ and $a_y(t) = [a_{y,1}(t), \cdots, a_{y,|K_y|}(t)]^T \in \mathbb{R}^{|K_y|}$. For each $i \in \{1, \cdots, |K_u|\}$, $a_{u,i}(t) \equiv 0$ for $t \in \mathbb{R}_{\geq 0}$ if no attack occurs on the ith transmission channel of $\mathcal{N}_a$, and similarly, for each $j \in \{1, \cdots, |K_y|\}$, $a_{y,j}(t) \equiv 0$ for $t \in \mathbb{R}_{\geq 0}$ if the jth transmission channel of $\mathcal{N}_s$ is not under attack. We suppose that the attacks occur at some unknown time instant T_0, and hence, $a_u(t) \equiv 0$ and $a_y(t) \equiv 0$ for $t < T_0$. In the sequel, the combined attack vector is denoted as $a(t) \triangleq [a_u^T(t), a_y^T(t)]^T \in \mathbb{R}^{|K_u|+|K_y|}$.

2.1 Closed-loop CPS

The closed-loop CPS including $\mathcal{C}$, $\mathcal{P}$, $\mathcal{N}_a$ and $\mathcal{N}_s$ are jointly denoted by Σ. Based on (1), in the presence of the integrity attack, Σ is described by

$$\Sigma : \begin{cases} \dot{x}(t) = Ax(t) + g(t, x) + Bu(t) + B_a a(t) \\ \qquad\quad +D_1 d(t), \\ u(t) = u_c(t, \tilde{y}(t), y_{\text{ref}}(t)) \\ y(t) = Cx(t) + D_2 d(t), \\ \tilde{y}(t) = y(t) + D_a a(t), \end{cases} \tag{2}$$

where $x \in \mathcal{X} \subset \mathbb{R}^{n_x}$ is the state vector ($\mathcal{X}$ is a compact subset of $\mathbb{R}^{n_x}$ containing the origin) and $y_{\text{ref}}(t) \in \mathbb{R}^{n_{\text{ref}}}$ is the output reference signal. In addition, $d(t) \in \mathbb{R}^{n_d}$ represents the lumped disturbances and noise, which is assumed to satisfy $\|d\|_{[t_1,t_2]} \leq \Delta$ for any $t_2 - t_1 \leq T$ where

$\Delta > 0$ is a scalar known by the defender and $T > 0$ is the evaluation time length. The matrices $A \in \mathbb{R}^{n_x \times n_x}$, $B \in \mathbb{R}^{n_x \times n_u}$, $C \in \mathbb{R}^{n_y \times n_x}$, $D_1 \in \mathbb{R}^{n_x \times n_d}$ and $D_2 \in \mathbb{R}^{n_y \times n_d}$, $B_a = [B\Gamma_u, 0_{n_x \times |K_y|}]$ and $D_a = [0_{n_y \times |K_u|}, \Gamma_y]$. In addition, the pair (A, C) is observable. The function $g : \mathbb{R}_{\geq 0} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$ represents the nonlinearity of the physical plant, which is known by the defender. The function $u_c : \mathbb{R}_{\geq 0} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_{\text{ref}}} \to \mathbb{R}^{n_u}$ is an output feedback control law (static or dynamic) such that in the nominal case ($d(t) \equiv 0$ and $a(t) \equiv 0$ for $t \in \mathbb{R}_{\geq 0}$), the closed-loop system can track y_{ref} asymptotically. Moreover, g and u_c satisfy $g(t, 0) = 0$ and $u_c(t, 0, 0) = 0$, $g(t, x)$ is piecewise continuous with respect to (w.r.t.) t and continuously differentiable w.r.t. x.

2.2 Anomaly Detector

We consider an anomaly detector $\mathcal{D}$ equipped with a residual generator $r(t)$ and a constant threshold J_{th}. Without loss of generality, it is supposed that the residual has a form as in Chen and Patton (1999) and is given as follows:

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

where $F : \mathbb{R}^{n_u} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_{\text{ref}}} \to \mathbb{R}^{n_y}$ is a model-based estimator to generate an estimate of $\tilde{y}(t)$. In the model-based fault diagnosis literature, such as Chen and Patton (1999); Blanke, Kinnaert, Lunze, Staroswiecki, and Schröder (2006); Ding (2013), model-based observers are frequently used as such an estimator. With the evaluation function $J(t) = \|r(t)\|_{\text{RMS}}$, the occurrence of an attack is ascertained if at some time $T_d > T_0$, the evaluation function $J(t)$ exceeds the constant threshold J_{th}, i.e.,

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

Since the residual $r(t)$ after the attack occurrence time T_0 (i.e., for $t > T_0$) is used to detect attacks, we refer to $\mathcal{D}$ as a forward-in-time detector in this paper. The majority of fault detectors in the literature such as in the books Chen and Patton (1999); Blanke et al. (2006); Ding (2013), are forward-in-time detectors. However, stealthy integrity attacks can result in residuals with sufficiently small amplitude, thereby passing such forward-in-time detectors without being detected. Consequently, it is necessary to develop new methodologies to detect stealthy integrity attacks. It is worth pointing out that there is no special requirement for the anomaly detector $\mathcal{D}$. Hence, $\mathcal{D}$ can be any anomaly detector with a residual $r(t)$ and a constant threshold J_{th} such that $J(t) \leq J_{th}$ in the nominal case (no anomalies).

In order to distinguish the variables, the superscript n is used in the normal case (attack free), while the superscript a is used to denote the changes of the variables due to attacks. For example, x^n is the plant state in the normal case and x^a is the change of x^n due to an attack, i.e., $x^a \triangleq x - x^n$.

2.3 Stealthy Integrity Attacks

Let us consider the transient processes of the integrity attacks such as replay attacks in Mo and Sinopoli (2009), covert attacks in Barboni et al. (2020) and zero-dynamics attacks in Teixeira et al. (2012, 2015a). The change of $\tilde{y}$ due to a replay attack may converge asymptotically to zero. In the presence of a covert attack, the change of $\tilde{y}$ also converges asymptotically to zero if the initial condition of the covert agent is nonzero. In the zero-dynamics attack case, if the non-exact values of the states of the physical plant are used by the attacker, then the value of $\tilde{y}$ may change instantaneously at the attack initiating time and then, such a change may go to zero asymptotically. Therefore, in some attack scenarios, the aforementioned attacks belong to a class of *undetectable attacks* as defined in Pasqualetti et al. (2013), but are not *perfectly undetectable attacks* as defined in Milošević et al. (2020). The transient process of the system outputs in the presence of the aforementioned class of undetectable attacks, in the context of asymptotically stable closed-loop CPS, is characterized in the following definition.

Definition 1. An integrity attack $a(t)$ initiated at time T_0, i.e., $a(t) \not\equiv 0$ for $t \geq T_0$, is considered as stealthy with respect to the anomaly detector $\mathcal{D}$ if

(a) $\|\tilde{y}(t) - \tilde{y}^n(t)\| \to 0$ as $t \to +\infty$;
(b) $0 < \|\tilde{y}(t) - \tilde{y}^n(t)\|_{\text{RMS}} \leq \delta$ for $t \geq T_0$ where δ is a sufficiently small scalar so that $J(t) \leq J_{th}$. ■

Remark 1. Condition (a) describes the asymptotic convergence of the transient process of the system outputs in the presence of the class of aforementioned undetectable attacks. Such a condition may be a result of the asymptotic stability of the closed-loop CPS, which is usually overseen in the related literature since the controllers are usually not taken into consideration. Condition (b) limits the maximum amplitude of the increments of the outputs under the attack such that it remains stealthy with respect to the anomaly detector $\mathcal{D}$. ∇

Remark 2. Undetectable attacks in Pasqualetti et al. (2013), perfectly undetectable attacks in Milošević et al. (2020) and stealthy integrity attacks defined in Definition 1 satisfy the inclusion-exclusion relation depicted in Fig. 2. Both perfectly undetectable attacks and stealthy integrity attacks belong to undetectable attacks, but

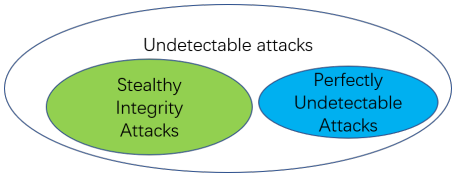

Fig. 2. Inclusion-exclusion relation among undetectable attacks, perfectly undetectable attacks and stealthy integrity attacks.

perfectly undetectable attacks and stealthy integrity attacks are independent and have no intersection. It should be noted that more disruption resources (i.e., larger number of disruption resources $|K_u| + |K_y|$, see Milošević et al. (2020)), are required by an attacker to launch perfectly undetectable attacks than in the case of the stealthy integrity attacks according to Definition 1 in this paper. For example, the disruption resources including all the sensor and actuator communication channels, i.e., $|K_u|+|K_y| = n_u+n_y$, are required by the perfectly undetectable covert attacks in Smith (2011). This indicates that the stealthy integrity attacks according to Definition 1 are easier to be realized and more practical for attackers in real-life applications. Hence, this paper considers only the case of stealthy integrity attacks defined in Definition 1 rather than the class of perfectly undetectable attacks. ∇

Remark 3. The stealthy integrity attacks satisfying Definition 1 are able to drive the system out of a safe region, and hence possess the same attack capability to compromise CPS as the zero-dynamics attacks in Teixeira et al. (2015a). For more details, the interested reader is referred to Teixeira et al. (2012, 2015a); Smith (2011); Barboni et al. (2020). ∇

Output zeroing strategies are widely used for generating integrity attacks for linear CPS such as Teixeira et al. (2012); Weerakkody et al. (2017). Such a strategy is also proved in Zhang et al. (2020) to be effective for the class of nonlinear CPS considered in this paper.In the sequel, we briefly present the output zeroing strategy in a system splitting manner.

1) Output zeroing strategy. By splitting the state x and the output $\tilde{y}(t)$ in (2) into $x = x_1 + z$ and $\tilde{y} = \tilde{y}_1 + \tilde{y}_2$, the system Σ in (2) can be split into Σ_1 and Σ_2 for $t \geq T_0$ where

$$\Sigma_1 : \begin{cases} \dot{x}_1(t) = Ax_1(t) + B_a a(t), \\ \tilde{y}_1(t) = Cx_1(t) + D_a a(t),\ x_1(T_0) = -z_0, \end{cases} \tag{5}$$

$$\Sigma_2 : \begin{cases} \dot{z}(t) = Az(t) + g(t,x) + Bu(t) + D_1 d(t), \\ \tilde{y}_2(t) = Cz(t) + D_2 d(t),\ z(T_0) = x(T_0) + z_0, \end{cases} \tag{6}$$

where $z_0 \in \mathbb{R}^{n_x}$ is a constant nonzero vector determined by the attack signal $a(t)$. It should be noted that $x_1(T_0)$ and $z(T_0)$ are chosen such that $x_1(T_0) + z(T_0) = x(T_0)$. Thus, the stealthy attack strategy for the nonlinear CPS (2) is proposed as follows:

$$\tilde{y}_1(t) = Cx_1(t) + D_a a(t) = 0,\ x_1(T_0) = -z_0 \neq 0, \tag{7a}$$
$$g(t, x_1 + z) = g(t, z),\ \forall\ t \geq T_0, \tag{7b}$$

where the initial condition z_0 is the *state-zero direction* and is discussed in detail later. The equation (7a) characterizes the output zeroing strategy for linear CPS in Pasqualetti et al. (2013) and (7b) is introduced such that $x_1(t)$ is decoupled with $g(t, x)$, thereby avoiding the negative effects of the nonlinear function $g(t, x)$ on the system output $\tilde{y}$ in the presence of the attacks. A nontrivial $a(t)$ satisfying (7a) exists if and only if the system Σ_1, with any control input matrices B_a and D_a, is not strongly observable [1]. Moreover, $x_1(t)$ must belong to the weakly unobservable subspace $\mathcal{V}(\Sigma_1)$, i.e., $x_1(t) \in \mathcal{V}(\Sigma_1)$ for $t \geq T_0$. In addition, according to the extended differential mean value theorem in Zemouche et al. (2005), we have $g(t, x_1 + z) - g(t, z) = \frac{\partial g}{\partial x} x_1(t)$. Thus, to guarantee (7b), $x_1(t)$ must belong to the kernel subspace $\mathcal{K}(g) = \ker(\frac{\partial g}{\partial x})$, i.e., $x_1(t) \in \mathcal{K}(g)$ for $t \geq T_0$. Instead of using the time-varying kernel subspace $\mathcal{K}(g)$, a linear time-invariant subspace $\mathcal{H}$ of $\mathcal{K}(g)$ is introduced. Hence, to guarantee (7b), one way is to guarantee that $x_1(t)$ belongs to the largest controlled invariant subspace $\mathcal{V}_{\mathcal{H}}(\Sigma_1)$ of Σ_1 contained in $\mathcal{H}$, i.e., $x_1(t) \in \mathcal{V}_{\mathcal{H}}(\Sigma_1)$ for $t \geq T_0$. Many types of nonlinear function $g(t, x)$ can satisfy (7b). Two intuitive examples are given in the following: 1) $g(t, x) = [x_2^2, x_2 \sin(x_2)]^T$ with $x = [x_1, x_2]^T$. In this example, $g(t, x)$ is independent of x_1 and $\mathcal{H} = \mathcal{K}(g) = \mathrm{Im}[1, 0]^T$; 2) $g(t, x) = [x_1 x_3 + x_2 x_3, x_1 x_4 + x_2 x_4, x_3 x_4, \sin(x_3 x_4)]^T$ where $x = [x_1, x_2, x_3, x_4]^T$. In this example, $\frac{\partial g(t,x)}{\partial x_1} = \frac{\partial g(t,x)}{\partial x_2}$ and $\mathcal{H} = \mathcal{K}(g) = \mathrm{Im}[1, -1, 0, 0]^T$.

The *state-zero* direction z_0 is proposed to satisfy

$$z_0 \in \mathcal{V}_0 \triangleq \mathcal{V}(\Sigma_1) \cap \mathcal{V}_{\mathcal{H}}(\Sigma_1),\ 0 < \|z_0\| \leq \delta_0, \tag{8}$$

where $\delta_0 > 0$ is a sufficiently small scalar. Using such a nontrivial z_0, the state of Σ_2 jumps immediately from $x(T_0)$ to $z(T_0) = x(T_0) + z_0$ when an attack satisfying (7a) and (7b) is initiated at T_0. Hence, $\tilde{y}(t) = \tilde{y}_2(t)$ also has a jump at T_0, allowing to detect the imposed stealthy integrity attacks.

Remark 4. A specific attack model satisfying the strategy (7a) and (7b) is given as follows:

$$a(t) = F_a \xi(t),\ \dot{\xi}(t) = (A + B_a F_a)\xi(t),\ \xi(T_0) = z_0,$$

where z_0 satisfies (8) and F_a satisfies $(A + B_a F_a)\mathcal{V}_0 \subset \mathcal{V}_0$, $(C + D_a F_a)\mathcal{V}_0 = 0$. Note that the attack model can be considered as an extension of the one generating zero-dynamics attacks. Similar to the requirements for generating zero-dynamics attacks, the initial state z_0 in the above attack model satisfies some geometric restrictions (see (8)) such that the system outputs are not changed (or slightly changed) in the presence of the generated attacks. In addition, it is worth pointing out that $\xi(T_0) = z_0$ is the only excitation resource for the attack model and thus, z_0 must be nonzero. ∇

[1] Trentelman et al. (2012)The system Σ_1 is strongly observable if for all $z_0 \in \mathbb{R}^{n_x}$ and any input $a(t)$, $y_1(t) = 0$ for $t \geq T_0$ implies $z_0 = 0$.

In order to guarantee (7a) and (7b) simultaneously, we have the following assumption.

Assumption 1. The attacker has the following *model knowledge*: the matrices A, $B\Gamma_u$ (or B_a), C and a linear subspace $\mathcal{H}$ of the kernel subspace $\mathcal{K}(g)$; and the following *disruption resources*: the communication channels $K_u \subseteq \{1, \cdots, n_u\}$ and $K_y \subseteq \{1, \cdots, n_y\}$, such that an $x_1(t) \not\equiv 0$ satisfying (7a) and (7b) exists. Moreover, the *state-zero direction* z_0 is chosen by the attacker to satisfy (8). ▼

Remark 5. The existence of an $x_1(t) \not\equiv 0$ satisfying (7a) and (7b) in Assumption 1 is equivalent to the existence of an attack $a(t) \not\equiv 0$ such that (7a) and (7b) hold, which can be guaranteed by $\mathcal{V}_0 = \mathcal{V}(\Sigma_1) \cap \mathcal{V}_{\mathcal{H}}(\Sigma_1) \neq \emptyset$. The model knowledge in terms of $g(t, x)$ and disruption resources K_u and K_y determine $\mathcal{V}_{\mathcal{H}}(\Sigma_1)$ and $\mathcal{V}(\Sigma_1)$ respectively. Therefore, the attacker should have sufficient *model knowledge* and the *disruption resources* such that $\mathcal{V}(\Sigma_1) \cap \mathcal{V}_{\mathcal{H}}(\Sigma_1) \neq \emptyset$. In addition, it is worth pointing out that even in the absence of Assumption 1, if the incremental system due to an attack can be expressed in the specific form (10) shown in the sequel, then the attack detection methodology developed in this paper can still detect it. Developing more relaxed conditions than the conditions (7a) and (7b) in Assumption 1 for generating stealthy integrity attack constitutes one of our future works. ▽

2) Incremental system. In the sequel, the incremental system of Σ due to an attack satisfying (7a) and (7b) will be derived. Let x^a, $\tilde{y}^a$ and u^a be the changes of x, $\tilde{y}$ and u respectively due to the attack, i.e., $x^a \triangleq x - x^n$, $\tilde{y}^a \triangleq \tilde{y} - \tilde{y}^n$ and $u^a \triangleq u - u^n$. According to the splitting of the system Σ shown in (5) and (6), the incremental system Σ^a of Σ can be split into the incremental systems Σ_1^a of Σ_1 and Σ_2^a of Σ_2, i.e., x^a and $\tilde{y}^a$ are split into $x^a(t) = x_1^a(t) + z^a(t)$ and $\tilde{y}^a(t) = \tilde{y}_1^a(t) + \tilde{y}_2^a(t)$ respectively, where

$$\Sigma_1^a : \begin{cases} \dot{x}_1^a(t) = Ax_1^a(t) + B_a a(t), \\ \tilde{y}_1^a(t) = Cx_1^a(t) + D_a a(t), \ x_1^a(T_0) = -z_0, \end{cases} \tag{9}$$

$$\Sigma_2^a : \begin{cases} \dot{z}^a(t) = \zeta(t, z^a), \\ \tilde{y}_2^a(t) = Cz^a(t), \ z^a(T_0) = z_0, \end{cases} \tag{10}$$

where $\zeta(t, z^a) \triangleq Az^a(t) + g(t, z) - g(t, z^n) + Bu^a(t) = Az^a(t) + \frac{\partial g}{\partial z} z^a(t) + Bu^a(t)$(see footnote [2]). Note that the systems (5) and (9) have the same dynamics and initial conditions. Thus, under the condition (7a), we can obtain that $\tilde{y}_1^a(t) = 0$ for $t \geq T_0$. Hence, $\tilde{y}^a(t) = \tilde{y}_2^a(t)$ and in the presence of a stealthy integrity attack satisfying (7a) and (7b), the increment $\tilde{y}_a(t)$ can be described by the output $\tilde{y}_2^a(t)$ of Σ_2^a in (10). Therefore, to satisfy condition (a) of Definition 1, the following additional restriction on the incremental system (10) is required.

Assumption 2. The system Σ_2^a in (10) is uniformly asymptotically stable for $z^a \in \mathcal{X}$. ▼

Remark 6. Assumption 2 indicates that $\tilde{y}^a(t)$ converges to zero asymptotically and hence, condition (a) in Definition 1 is satisfied. In fact, Assumption 2 is an implicit stability restriction on the system Σ in (2). Based on the equivalence between asymptotic stability and incrementally asymptotic stability given in Angeli (2002), a sufficient condition to guarantee Assumption 2 is that in the nominal case and for $y_{\text{ref}}(t) \equiv 0$ for $t \geq 0$, Σ is asymptotically stable. Regarding the incremental stability, the interested reader is referred to Angeli (2002). ▽

Remark 7. Similar to the attack strategy in Pasqualetti et al. (2013, 2015), the attack strategy (7) is able to generate replay attacks, zero-dynamics attacks and covert attacks in the scenarios that these attacks are stealthy but not completely stealthy (see the stealthy integrity attacks in Fig. 2). It is worth pointing out that in the scenarios where the attacker can access only part of the actuator and sensor communication channels, (i.e., $|K_u| < n_u$ and $|K_y| < n_y$), the covert attack signals $a_u(t)$ and $a_y(t)$ may need to be specifically designed using the attack strategy (7) with a nonzero z_0. Regarding the specific conditions on K_u and K_y that are required, so that the covert attack signals a_u and a_y can be generated by the attack strategy (7) with a nonzero z_0, this will be dealt with in future work. In the presence of the aforementioned replay attacks, zero-dynamics attacks and covert attacks, the incremental system can be characterized by (10) with a nonzero z_0. Hence, the developed methodology presented in the sequel is effective in detecting replay attacks and zero-dynamics attacks, and is also able to detect covert attacks in the case that the covert attack signals a_u and a_y are generated by the attack strategy (7) with a nonzero z_0. In addition, for the stealth multiplicative attacks in Na and Eun (2018), since such attacks are completely stealthy, the detection methodology developed in this paper may not be able to detect them. Additional approaches such as active detection methodologies should be exploited to detect these attacks, which is a future research direction. ▽

In the rest of this paper, "*stealthy integrity attacks*" is refer to the attacks satisfying Definition 1 and generated based on the output zeroing strategy characterized by (7a), (7b) and (8). Moreover, in the following sections, the incremental system Σ_2^a satisfying Assumption 2 will be exploited to detect the stealthy integrity attacks.

[2] In Σ_2^a, $g(t, z) - g(t, z^n)$ is replaced with $\frac{\partial g}{\partial z} z^a(t)$ since $g(t, z) - g(t, z^n) = \frac{\partial g}{\partial z} z^a(t)$ where

$$\frac{\partial g}{\partial z} \triangleq \begin{bmatrix} \frac{\partial g_1}{\partial z_1} & \cdots & \frac{\partial g_1}{\partial z_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial z_1} & \cdots & \frac{\partial g_n}{\partial z_n} \end{bmatrix},$$

with g_i being the ith element of g, and z_i being the ith element of z.

3 Feasibility Analysis of Equivalent Change to Trigger Alarms

The detection scheme to be developed is a backward-in-time detector. Backward-in-time detectors are different from forward-in-time detectors in terms of the time instant at which the system states are estimated. Traditional forward-in-time detectors estimate the system states at the time instant posterior to the attack occurrence time, while backward-in-time detectors estimate the system states at the time instant prior to the attack occurrence time. Backward-in-time detectors are inherently more suitable for dealing with the stealthiness problems. Intuitively, when running reversely in time, a stable system with a system matrix A becomes an unstable amplifier with $-A$. A *change* of the system state occurs after an attack happens. The *equivalent change* (mathematically defined later) at a time prior to the attack occurrence time, is recovered by this stable system running reversely in time, and thus, is an amplified quantity of the *change*. Backward-in-time detectors use this *equivalent change* for detecting stealthy integrity attacks that otherwise would remain undetected. Next, we investigate the *equivalent change* to show that although the true *change* can not trigger any alarms in an attack case, the *equivalent change* can.

The solution to the differential system (10) can be described by the time-dependent flow of the vector field ζ. In particular, the solution at any time $t + T_0$ for $t \in \mathbb{R}$ starting at time T_0 with initial condition $z^a(T_0)$, which is collectively defined as $(T_0, z^a(T_0))$, can be described by $z^a(t + T_0) = \psi(t,\ T_0,\ z^a(T_0))$. Such a solution satisfies the following equation, that is for any $\tau,\ t \in \mathbb{R}$,

$$\frac{d\psi(t,\ T_0,\ z^a(T_0))}{dt} = \zeta\left(t + T_0, \psi(t,\ T_0,\ z^a(T_0))\right), \quad (11)$$

$$\begin{aligned} z^a(\tau + t + T_0) &= \psi(\tau,\ t + T_0,\ \psi(t,\ T_0,\ z^a(T_0)) \\ &= \psi(\tau + t,\ T_0,\ z^a(T_0)). \end{aligned} \quad (12)$$

In addition, if the initial condition $z^a(T_0) = 0$, then the solution $z^a(t) = 0$ identically for $t \geq T_0$, i.e.,

$$\psi(t - T_0,\ T_0,\ 0) = 0,\ \forall\ t \in \mathbb{R}. \quad (13)$$

More details about the flows of vector fields can be found in Isidori (2013). The equation (12) explicitly indicates that for any time instant $\tau + t + T_0 < T_0$, the solution $z^a(\tau + t + T_0)$ does exist. Also, based on (11), for two time instants t_b and t satisfying $t_b < t$, $z^a(t_b)$ and $z^a(t)$ have the following relation:

$$z^a(t_b) = \psi(t_b - t,\ t,\ z^a(t)), \quad (14)$$

$$z^a(t) = \psi(t - t_b,\ t_b,\ z^a(t_b)). \quad (15)$$

Subsequently, based on (14), the equivalent quantity of z^a at a time t_b is defined as follows.

Definition 2. For the state $z^a(t)$ of the system (10) starting at $(T_0, z^a(T_0))$, its equivalent quantity at the time t_b where $t_b < T_0 \leq t$, recovered based on $z^a(t)$, is defined by

$$z^a(t_b|t) \triangleq \psi(t_b - t,\ t,\ z^a(t)), \quad (16)$$

where $z^a(t_b|t)$ denotes the equivalent quantity of z^a at t_b recovered based on the quantity of $z^a(t)$. ■

To intuitively explain the above definition, we consider a special example that the system (10) is a linear system, i.e., $\dot{z}^a = Az^a$. In this case, $z^a(t_b|t)$ in (16) can be written as $z^a(t_b|t) = \exp(A(t_b - t))z^a(t) = \exp(-A(t - t_b))z^a(t)$ where $t_b < t$. This example shows that when time runs backwards, a stable system with a system matrix A becomes an unstable amplifier with system matrix $-A$. The properties of the equivalent quantity $z^a(t_b|t)$ in (16) are summarized in the following lemma. To this end, based on Assumption 2, $z^a(t)$ can be bounded as follows

$$\|z^a(t)\| \leq \beta\left(\|z_0\|, t - T_0\right), \forall\ t \geq T_0, \quad (17)$$

where β is a $\mathcal{KL}$ class function deduced based on Assumption 2.

Lemma 1. *The equivalent quantity $z^a(t_b|t)$ in (16) has the following properties:*
(a) *For any time $t \geq T_0$, a fixed time t_b such that $t_b < T_0 \leq t$, and any $z_0 \in \mathbb{R}^{n_x}$, the quantity $z^a(t_b|t)$ is a constant vector with respect to the time t, i.e.,*

$$\frac{d\ z^a(t_b|t)}{d\ t} = 0,\ \forall\ t \geq T_0. \quad (18)$$

(b) *In the non-attack case, $z^a(t_b|t)$ satisfies*

$$\frac{d\ z^a(t_b|t)}{d\ t} = 0,\ z^a(t_b|t) = 0,\ \forall\ t < T_0. \quad (19)$$

(c) *In the presence of a stealthy integrity attack at T_0, if $T_0 \in [t_b, t_s]$ where t_s is a fixed time instant, then for a fixed time t_b, there exists a $\mathcal{K}$ class function ρ_{t_s} such that*

$$\|z^a(t_b|t)\| \leq \rho_{t_s}^{-1}(\delta_0),\ \forall\ t \geq T_0, \quad (20)$$

where the function $\rho_{t_s}^{-1}$ represents the inverse of the function ρ_{t_s} defined by

$$\rho_{t_s}(\delta_0) \triangleq \beta(\delta_0, t_s - t_b),\ \forall\ \delta_0 \geq 0, \quad (21)$$

with δ_0 being given in (8), and β being given in (17). ■

Proof. **(a)** By using (12) for $\tau = -T_0$, we have

$$z^a(t) = \psi(t - T_0,\ T_0,\ z^a(T_0)).$$

Then, by using (14), $z^a(t_b|t)$ defined by (16) can be written as

$$\begin{aligned} z^a(t_b|t) &= \psi(t_b - t,\ t,\ z^a(t)) \\ &= \psi(t_b - t,\ t,\ \psi(t - T_0,\ T_0,\ z^a(T_0))). \end{aligned}$$

By using the corresponding relation between the first line and the second line in (12), we obtain that

$$z^a(t_b|t) = \psi(t_b - T_0,\ T_0,\ z^a(T_0)). \tag{22}$$

Since both t_b and T_0 are fixed time instants, and z_0 is also fixed, $\psi(t_b - T_0,\ T_0,\ z_0)$ is fixed, hence $z^a(t_b|t)$ is a constant vector, and (18) follows.
(b) In the non-attack case, no change happens, i.e., $z^a(t) = 0$, then $z^a(T_0) = 0$ (in this case, T_0 can be any time instant). Then it follows from (22) and (13) that $z^a(t_b|t) = \psi(t_b - T_0,\ T_0,\ 0) = 0$.
(c) From (17) and $z^a(T_0) = z_0$, $z^a(t_b)$ satisfies

$$\|z_0\| = \|z^a(T_0)\| \leq \beta(\|z^a(t_b|t)\|, T_0 - t_b),\ \forall\ t \geq t_b.$$

Let $\rho_{T_0}(x)$ be a $\mathcal{K}$ class function defined by $\rho_{T_0}(x) \triangleq \beta(x, T_0 - t_b)),\ \forall\ x \geq 0$, and $\rho_{T_0}^{-1}(x)$ be its inverse function (note that $\rho_{T_0}^{-1}(x)$ is also a $\mathcal{K}$ class function). Then, we can derive $\|z_0\| = \|z^a(T_0)\| \leq \rho_{T_0}(\|z^a(t_b|t)\|)$, and further, for z_0 satisfying (8), we have

$$\|z^a(t_b|t)\| \leq \rho_{T_0}^{-1}(\|z_0\|) \leq \rho_{T_0}^{-1}(\delta_0).$$

Since for the scalar δ_0, $\beta(\delta_0, t_s - t_b) \leq \beta(\delta_0, T_0 - t_b)$, then $\rho_{t_s}^{-1}(\delta_0) \geq \rho_{T_0}^{-1}(\delta_0)$. Thus, we have

$$\|z^a(t_b|t)\| \leq \rho_{T_0}^{-1}(\|z_0\|) \leq \rho_{t_s}^{-1}(\delta_0).$$

Hence, the inequality (20) follows. □

Remark 8. Result (a) in Lemma 1 indicates that for a fixed $t_b < T_0 < t$, $z^a(t_b|t)$ is a constant vector as $z^a(t)$ converges to zero asymptotically. Intuitively, $z^a(t_b|t)$ is the analytical value of $z^a(t)$ at the fixed time instant t_b, which is constant, since $z^a(t)$ is the solution of the differential equation in (10) and its value at the fixed time instant t_b is constant. The residual $r(t)$ of the detector $\mathcal{D}$ in (3) converges to zero if the increment $z^a(t)$ goes to zero asymptotically, which is one reason that typical forward-in-time detectors such as $\mathcal{D}$ can not detect stealthy integrity attacks. The property of $z^a(t_b|t)$ shown in result (a) implies that $z^a(t_b|t)$ provides an effective way to solve the aforementioned problem of forward-in-time detectors in detecting stealthy integrity attacks. A well designed residual based on $z^a(t_b|t)$ can retain a constant vector value as $z^a(t)$ converges to zero asymptotically, thereby providing the possibility to detect stealthy integrity attacks. In addition, it should be noted that the control law u_c affects the function ρ_{t_s} in (21). Since the control law u_c can affect the dynamics of the system Σ_2^a in (10) through u^a, it can also influence the value of $z^a(t_b|t)$ defined in (16) and the function β in (17). Thus, based on (21) in Lemma 1, u_c can affect the function ρ_{t_s}, which will then affect the estimation of $z^a(t_b|t)$ (see (34) in the next section). ∇

In the sequel, we will prove that $z^a(t_b|t)$ can be used to trigger alarms in the presence of an attack, while not being impacted in the non-attack case. We start by defining the new residual. By using $z^a(t_b|t)$, the equivalent change of $\tilde{y}^a(t)$ is given by

$$\tilde{y}^a(t_b|t) \triangleq C z^a(t_b|t). \tag{23}$$

Then, a new residual, referred to as backward-in-time residual and denoted by $r\,(t_b|t)$, is proposed as follows:

$$r\,(t_b|t) \triangleq r(t_b) + \tilde{y}^a(t_b|t). \tag{24}$$

The feasibility theorem is presented in the following.

Theorem 1. *For the system (10), the equivalent quantity $z^a(t_b|t)$ in (16) and the residual $r\,(t_b|t)$ in (24) satisfy the following boundedness properties regarding the stealthy attacks considered in this paper:*
(a) *In the absence of an attack, the backward-in-time residual satisfies*

$$\|r\,(t_b|t)\,\|_{\text{RMS}} \leq J_{th},\ \forall\ t_b < T_0 \leq t, \tag{25}$$

where J_{th} is the threshold given in (4).
(b) *In the presence of a stealthy integrity attack, and under Assumption 2, there exists a time instant $t_b < T_0$ such that the backward-in-time residual satisfies*

$$\|r(t_b|t)\|_{\text{RMS}} > J_{th},\ \forall\ t \geq T_0. \tag{26}$$

Moreover, for the fixed times t_b and $T_0 \in [t_b, t_s]$, and z_0 satisfying (8), $z^a(t_b|t)$ is bounded by

$$\frac{J_{th} + \|r(t_b)\|}{\|C\|} < \|z^a(t_b|t)\| \leq \rho_{t_s}^{-1}(\delta_0), \tag{27}$$

where δ_0 is given in (8). ■

Proof. **(a)** Based on Lemma 1, in the absence of an attack, $z^a(t_b|t) = 0$. From (23), $\tilde{y}^a(t_b|t) = 0$ and thus, $r\,(t_b|t) = r(t_b)$. Since at time $t_b < T_0$, no attack is present, $\|r(t_b)\|_{\text{RMS}} \leq J_{th}$ and inequality (25) follows.
(b) Based on the definition of uniformly asymptotic stability in Khalil (2001), given the system (10) satisfying Assumption 2, we can deduce that for any $\eta > \sup_{t \geq T_0} \|z^a(t)\|$, there always exists a $T_1 = T_1(\eta) > 0$ such that

$$\exists\ t_b \leq T_0 - T_1,\ \|z^a(t_b|t)\| > \eta,\ \forall\ \|z^a(t)\| \leq \eta.$$

Since $\eta > \sup_{t\geq T_0} \|z^a(t)\|$, then a sufficiently large η exists that additionally satisfies $\|C\|\eta \geq J_{th} + \|r(t_b)\|$ where $J_{th} + \|r(t_b)\| > \|C\| \sup_{t\geq T_0} \|z^a(t)\|$ due to the stealthiness of the attack. Thus, we can obtain

$$\exists\, t_b \leq T_0 - T_1,\ \|\tilde{y}^a(t_b|t)\| > J_{th} + \|r(t_b)\|,\ \forall\, \|z^a(t)\| \leq \eta. \tag{28}$$

where $\tilde{y}^a(t_b|t)$ is defined in (23). Note that based on the result (a) in Lemma 1 and (23), $\tilde{y}^a(t_b|t)$ is a constant vector, and from (24), $r(t_b|t)$ is a constant vector as well. Therefore, $\|\tilde{y}^a(t_b|t)\| = \|\tilde{y}^a(t_b|t)\|_{\text{RMS}}$, $\|r(t_b)\| = \|r(t_b)\|_{\text{RMS}}$ and $\|r(t_b|t)\| = \|r(t_b|t)\|_{\text{RMS}}$. Thus, from (24) and (28), and using the triangle inequality of vector norms, we have

$$\|r(t_b|t)\|_{\text{RMS}} \geq \|\tilde{y}^a(t_b|t)\|_{\text{RMS}} - \|r(t_b)\|_{\text{RMS}} > J_{th}.$$

Hence, the inequality (26) follows. Moreover, from $\|\tilde{y}^a(t_b|t)\| = \|Cz^a(t_b|t)\| > J_{th} + \|r(t_b)\|$, the left hand side of (27) follows. The right hand side follows directly from (20) in Lemma 1. □

From Theorem 1, we conclude that *equivalent changes* are able to trigger alarms in the presence of *stealthy integrity attacks*. However, it should be noted that the *equivalent change* $z^a(t_b|t)$ is not available to the defender. One reason is that it may be very difficult or impossible to explicitly find the analytical solution to ordinary differential equation (11) due to the gradient of $g(t, z)$ in (10) (see $\zeta(t, z^a)$ below (10)). Another reason is that $z^a(t)$ is not available to the defender. Therefore, a task of the backward-in-time detector is to construct a procedure for estimating the *equivalent changes*. In the following, an optimal fixed-point smoother will be designed to estimate the *equivalent change* $z^a(t_b|t)$.

4 $\mathcal{H}_\infty$ Fixed-point Smoothing Scheme

A fixed-point smoother provides an on-line backward-in-time estimation procedure, yielding an estimate of a signal at the current time instant using past and current observed measurements at the first stage and then updating it using the observed measurements as time progresses. Therefore, fixed-point smoothing provides a useful tool for estimating the *equivalent change* at a fixed time. Fixed-point smoothers for linear systems has been well studied in Meditch (1967); Simon (2006); Einicke (2019) and references therein, and $\mathcal{H}_\infty$ fixed-point smoothers have been established in Shaked and Theodor (1992); Theodor and Shaked (1994); Einicke (2019). However, $\mathcal{H}_\infty$ fixed-point smoothers for nonlinear systems have not been well studied, and to the authors' best knowledge, fixed-point smoothers have not been used for detecting malicious cyber attacks. In the sequel, a particular $\mathcal{H}_\infty$ fixed-point smoother for a class of Lipschitz nonlinear systems is designed with the task to estimate the *equivalent change* $z^a(t_b|t)$.

To this end, some preliminaries are given. A finite-time horizon detection scheme operating in $[t_b, t_s]$ is designed in the sequel. Such a detection scheme is sequentially implemented (repeatedly with possible partial overlapping time intervals) in the practical application and thus, t_b and t_s are updated in each repetition. A simple example is given in Fig. 3 to show the implementation and update approach. It can be seen that in the 1st repetition, $t_{b,1} = t_1$ and $t_{s,1} = t_4$, in the second repetition t_b and t_s are updated by $t_{b,2} = t_2$ and $t_{s,2} = t_5$, and in the third application, $t_{b,3} = t_3$ and $t_{s,3} = t_6$. Moreover, the attack occurrence time T_0 is located between $t_{b,3}$ and $t_{s,3}$, i.e., $T_0 \in [t_{b,3}, t_{s,3}]$, and hence belongs to the third repetition. Both $t_{b,i}$ and $t_{s,i}$ for all repetitions can be set off-line by the designer. A simple way is to set the $t_{b,i}$ every T_s seconds with a sequential overlapping window of length $2T_s$ seconds. Then, the 1st repetition is conducted in the interval $[0, 2T_s]$, the second in $[T_s, 3T_s]$ and the third in $[2T_s, 4T_s]$. So in this example, at any given time instant after T_s, two repetitions run in parallel. In

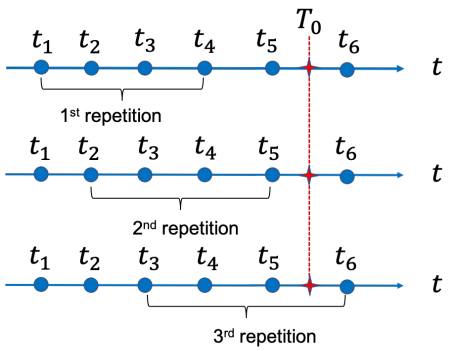

Fig. 3. Example of the sequential implementation for a finite–time horizon detection scheme and the update approach for t_b and t_s.

addition, the computation burden of the implementation approach in a finite time duration is determined by the amount of the repetitions during this time duration, which is affected by the length of the time duration between two repetitions (i.e., $t_{b,2} - t_{b,1}$ and $t_{b,3} - t_{b,2}$). In general, the scheme allows one or more repetitions to be active at any given time, which is determined by the designer. Therefore, in the practical application of the sequential implementation approach, the computation burden should be considered in determining the implementation frequency of the repetitions.

Hence, by using such a sequential implementation approach, we can consider that a time interval $[t_b, t_s]$ including the attack occurrence time instant T_0 and satisfying result (b) of Theorem 1 always exists. Therefore, it is reasonable to study only the case that the attack occurrence time satisfies

$$T_0 \in [t_b, t_s], \tag{29}$$

and consider that result (b) in Theorem 1 is satisfied

so that the attack is detectable in the context of the *equivalent change* at the time t_b.

Remark 9. In the proposed implementation approach, the detection scheme is repeatedly implemented in a reasonably short time after the start of the previous repetition, leading to partially overlapping repetition time interval, i.e., $t_{b,i+1} \in [t_{b,i}, t_{s,i}]$ and $t_{s,i+1} > t_{s,i}$, $i = 1, 2, 3, \cdots, N$ where $t_{b,i}$ and $t_{s,i}$ indicate the start time and the end time t_b and t_s of the ith repetition respectively. Such an approach facilitates the condition that the length between T_0 and t_b is sufficiently large such that result (b) in Theorem 1 can be guaranteed. Note that at any given time more than one repetition may be running (e.g., in Fig. 3, in the time interval $[t_3, t_4]$, three repetitions are active). ∇

In addition, we make the following assumption regarding the nonlinear function g.

Assumption 3. The function $g(t, x)$ is locally Lipschitz with respect to x, i.e.,

$$\|g(t, x) - g(t, \hat{x})\| \leq l\|x - \hat{x}\|, \ \forall \ x, \ \hat{x} \in \mathcal{X},$$

where l is the known Lipschitz constant. ▼

Remark 10. The Lipschitz condition in Assumption 3 is only needed for the fixed-point smoother design in the sequel. In the absence of the Lipschitz condition, the results in Lemma 1 and Theorem 1 still hold. Fixed-point smoother design for a more general class of nonlinear systems constitutes an aspect of our future works. ∇

4.1 Fixed-point Smoother Design

In this part, one of the repetition of the aforementioned sequential implementation is considered, and the time duration is denoted as $[t_b, t_s]$. In addition, we assume $t_s - t_b \leq T$ such that $\|d(t)\|_{[t_s, t_b]} \leq \Delta$. We start by designing a fixed-point smoother working in the finite-time interval $[t_b, t_s]$ for estimating $z^a(t_b|t)$. Following the state augmentation design approach of the fixed-point smoother in Simon (2006); Einicke (2019), a new state variable is introduced as follows:

$$\phi(t) \triangleq z^a(t_b|\tau), \tau \geq T_0, \ \forall \ t \geq t_b. \tag{30}$$

Since t_b satisfies result (b) in Theorem 1, then $\phi(t)$ satisfies the inequality (27). Moreover, since $z^a(t_b|\tau)$ satisfies (18) in Lemma 1 for $\tau \geq T_0$, then we have

$$\frac{d\ \phi(t)}{d\ t} = \dot{\phi}(t) = 0, \ \forall \ t \in [t_b, t_s]. \tag{31}$$

The smoother is designed based on the equivalent system Σ_2 given in (6) in the presence of an attack. It should be noted that, as analyzed after equation (10), Σ_2 can describe the dynamics of the system Σ in the attack scenario, i.e., $t \geq T_0$. Moreover, since Σ_2 and Σ with $a(t) = 0$ have the same dynamics, Σ_2 can also represent the dynamics of the system Σ in the attack-free time duration, i.e., for $t < T_0$. Therefore, referring to the structure of Σ_2, the fixed-point smoother over the time interval $[t_b, t_s]$ is designed as follows:

$$\mathcal{S}: \begin{cases} \dot{\hat{z}}(t) = A\hat{z}(t) + g\left(t, \hat{z}\right) + Bu(t) \\ \qquad\qquad -K_z(t)\left(\tilde{y}(t) - C\hat{z}(t)\right), \\ \dot{\hat{\phi}}(t) = -K_\phi(t)\left(\tilde{y}(t) - C\hat{z}(t)\right), \\ \hat{\tilde{y}}^a(t_b|t) = C\hat{\phi}(t), \end{cases} \tag{32}$$

where $\hat{z} \in \mathbb{R}^{n_x}$ is the estimate of the state z of the system Σ_2, $\hat{\phi} \in \mathbb{R}^{n_x}$ is the estimate of ϕ, and $\hat{\tilde{y}}^a(t_b|t) \in \mathbb{R}^{n_y}$ is the estimate of $\tilde{y}^a(t_b|t)$ in (23). Moreover, $K_z(t) \in \mathbb{R}^{n_x \times n_y}$ and $K_\phi(t) \in \mathbb{R}^{n_x \times n_y}$ are time-varying gain matrices to be optimized in the next subsection.

This smoother is activated at the time t_b and lasts until t_s. Different from the standard fixed-point smoother in Einicke (2019); Simon (2006), the initial conditions $\hat{\phi}(t_b)$ and $\hat{z}(t_b)$ should be specially designed. Specifically, the initial condition $\hat{\phi}(t_b)$ is chosen to be close to the true value of $\phi(t)$, i.e., satisfying the inequality (27), and in addition, $\hat{\phi}(t_b)$ must not trigger any alarm at $t = t_b$. In order not to trigger any alarm at t_b, based on the residual defined in (24), the following condition must be satisfied: $\|r(t_b) + \hat{\tilde{y}}^a(t_b|t_b)\| = \|r(t_b) + C\hat{\phi}(t_b)\| \leq \hat{J}_{th}$ where $\hat{J}_{th}$ is the new threshold to be determined later. Therefore, $\|\hat{\phi}(t_b)\| \leq \frac{\hat{J}_{th} - \|r(t_b)\|}{\|C\|}$ and by combining (27), we have $\frac{J_{th} + \|r(t_b)\|}{\|C\|} < \|\hat{\phi}(t_b)\| \leq \min\left\{\frac{\hat{J}_{th} - \|r(t_b)\|}{\|C\|}, \ \rho_{t_s}^{-1}(\delta_0)\right\}$, where δ_0 is given in (8). Based on the aforementioned selection of $\hat{\phi}(t_b)$, one can yield

$$\|\hat{\phi}(t_b) - \phi(t_b)\| \leq \delta_1, \tag{33}$$

where

$$\delta_1 \triangleq \min\left\{\frac{\hat{J}_{th} - \|r(t_b)\|}{\|C\|}, \ \rho_{t_s}^{-1}(\delta_0)\right\} + \rho_{t_s}^{-1}(\delta_0), \tag{34}$$

with δ_0 being given in (8). Moreover, the initial condition $\hat{z}(t_b)$ is chosen by considering the confidence in the knowledge of $z(t_b)$. In order to guarantee that $\hat{z}(t_b)$ and $\hat{\phi}(t_b)$ have the same confidence in the knowledge of $z(t_b)$ and $\phi(t_b)$ respectively (further explanation on this is given in the sequel), $\hat{z}(t_b)$ is also chosen to satisfy

$$\|z(t_b) - \hat{z}(t_b)\| \leq \delta_1. \tag{35}$$

Note that there is no attack at the time t_b (see (29)) and a well designed observer such as the observer of the anomaly detector $\mathcal{D}$ can provide an accurate estimate of

the state $z(t_b)$. Hence, $z(t_b)$ is considered as known by the defender in selecting $\hat{z}(t_b)$ satisfying (35).

4.2 Optimal Parameters Design

The optimization problem arises due to the presence of the disturbances in the system. To this end, a compact form of the estimation error system is given. Some notations are first introduced as follows:

$$\bar{A} \triangleq \mathrm{diag}(A, 0),\ \bar{C}_1 \triangleq [C,\ 0],\ \bar{C}_2 \triangleq [0,\ C], \quad (36)$$

$$\bar{g}^T(t, z, \hat{z}) \triangleq (g(t, z) - g(t, \hat{z}))^T [I_{n_x}, 0], \quad (37)$$

$$K^T(t) \triangleq [K_z^T(t), K_\phi^T(t)],\ \bar{D}_1 \triangleq [D_1^T, 0]^T. \quad (38)$$

By defining $e_z(t) \triangleq z(t) - \hat{z}(t)$ and $e_\phi(t) \triangleq \phi(t) - \hat{\phi}(t)$ as the estimation errors of z and ϕ respectively, a compact error is defined as follows: $e^T(t) \triangleq [e_z^T(t), e_\phi^T(t)]$. Moreover, let $e_y(t) \triangleq \tilde{y}^a(t_b|t) - \hat{\tilde{y}}^a(t_b|t)$ denote the estimation error of $\tilde{y}^a(t_b|t)$. Then, from (28), (31) and (32), the error system can be obtained as follows:

$$\mathcal{E}: \begin{cases} \dot{e}(t) = (\bar{A} + K(t)\bar{C}_1)e(t) + \bar{g}(t, z(t), \hat{z}(t)) \\ \qquad\quad +[\bar{D}_1 + K(t)D_2]d(t), \\ e_y(t) = \bar{C}_2 e(t). \end{cases} \quad (39)$$

Next, the estimation of the backward-in-time residual $r(t_b|t)$ in (24) is constructed. Based on (24), the estimation is proposed by using the estimated equivalent change $\hat{\tilde{y}}^a(t_b|t)$ provided by the smoother $\mathcal{S}$ as

$$\hat{r}(t_b|t) \triangleq r(t_b) + \hat{\tilde{y}}^a(t_b|t). \quad (40)$$

From $\hat{\tilde{y}}^a(t_b|t) = \tilde{y}^a(t_b|t) - e_y(t)$, $\hat{r}(t_b|t)$ can be split into

$$\begin{aligned} \hat{r}(t_b|t) &= r(t_b) + \tilde{y}^a(t_b|t) - e_y(t) \\ &= r(t_b|t) - e_y(t). \end{aligned} \quad (41)$$

Motivated by the optimal residual design methodology in Ding (2013), $\hat{r}(t_b|t)$ is to be optimized to achieve the $\mathcal{H}_\infty$ performance with respect to the disturbance $d(t)$. More specifically, according to (41), since $r(t_b|t)$ is fixed, the optimization problem is formulated as designing $K_z(t)$ and $K_\phi(t)$ to satisfy the $\mathcal{H}_\infty$ performance given as follows:

$$\frac{\|e_y(t)\|^2_{[t_b,t_s]}}{e^T(t_b)\Theta e(t_b) + \|d(t)\|^2_{[t_b,t_s]}} \le \gamma^2, \quad (42)$$

where $\gamma > 0$ is the $\mathcal{H}_\infty$ performance index and $\Theta = \Theta^T \ge 0$ has the following structure:

$$\Theta = \frac{1}{4}\begin{bmatrix} \Theta_0 & \Theta_0 \\ \Theta_0 & \Theta_0 \end{bmatrix}, \quad (43)$$

where $\Theta_0 = \Theta_0^T > 0$. It is worth pointing out that the special structure of Θ reflects the same confidence in the knowledge of the initial conditions $e_z(t_b)$ and $e_\phi(t_b)$, which is a result of the same bounds of $\phi(t_b) - \hat{\phi}(t_b)$ and $z(t_b) - \hat{z}(t_b)$ given in (33) and (35) respectively. The conditions for guaranteeing the $\mathcal{H}_\infty$ performance (42) are given in the following theorem.

Theorem 2. *Suppose that Assumption 3 holds. Then, for a given performance index $\gamma > 0$, the estimation error system $\mathcal{E}$ (39) satisfies the $\mathcal{H}_\infty$ performance (42) if there exists a solution $Q(t) = Q^T(t) \ge 0$ in the time interval $[t_b, t_s]$ to the following differential Riccati equation:*

$$\begin{aligned} \dot{Q} = &\left(\bar{A} - \bar{D}_1 D_2^T R^{-1}\bar{C}_1\right) Q + Q\left(\bar{A} - \bar{D}_1 D_2^T R^{-1}\bar{C}_1\right)^T \\ &- Q\left(\bar{C}_1^T R^{-1}\bar{C}_1 - \mu^{-2} l^2 E_{n_x} - \gamma^{-2}\bar{C}_2^T\bar{C}_2\right) Q \\ &+ \mu^2 E_{n_x} + \bar{D}_1\left(I - D_2^T R^{-1} D_2\right)\bar{D}_1^T, \end{aligned} \quad (44)$$

where μ is any positive scalar, $R = D_2 D_2^T$, $E_{n_x} = [I_{n_x}, 0]^T[I_{n_x}, 0]$ and

$$Q(t_b) = \begin{bmatrix} \Theta_0^{-1} & \Theta_0^{-1} \\ \Theta_0^{-1} & \Theta_0^{-1} \end{bmatrix}.$$

Then, $K_z(t)$ and $K_\phi(t)$ in (32) are obtained as

$$K_z(t) = [I_{n_x}, 0]K(t),\ K_\phi = [0, I_{n_x}]K(t), \quad (45)$$

where $K(t)$ is given by

$$K(t) = \left(Q\bar{C}_1^T + \bar{D}_1 D_2^T\right) R^{-1}. \quad (46)$$

■

Proof. See *Appendix: Proof of Theorem 2.* □

4.3 Residual Evaluation Function and Threshold Generation

In the context of fault diagnosis, detection residuals are evaluated to form evaluation functions (see, e.g., Ding (2013)). This paper also designs the evaluation function by evaluating the residual $\hat{r}(t_b|t)$ in (40). In the design of the residual evaluation function, the normalization problem should be considered. The residual evaluation is a function such that when there is no attack, the function output is close or equal to zero (similar to residual evaluations in fault diagnosis schemes, in the absence of faults). Based on this requirement, the following evaluation function is proposed:

$$\hat{J}(t_b|t) \triangleq \|\hat{r}(t_b|t)\|_{\mathrm{RMS}} - \|r(t_b) + C\hat{\phi}(t_b)\|_{\mathrm{RMS}}, \quad (47)$$

where the residual $\hat{r}\,(t_b|t)$ is defined in (40) is a time-varying function and is sensitive to attacks. The term $\|r(t_b) + C\hat{\phi}(t_b)\|_{\rm RMS}$ is included in $\hat{J}\,(t_b|t)$ since it constitutes a correction term (constant bias) required for normalization purposes, in order to guarantee that $\hat{J}\,(t_b|t)$ is close to zero in the absence of attacks. The following lemma is given to show the boundedness properties of $\hat{J}\,(t_b|t)$ in the absence of an attack.

Lemma 2. *In the absence of an attack, the evaluation function $\hat{J}\,(t_b|t)$ in (47) satisfies*

$$\hat{J}\,(t_b|t) \leq \hat{J}_{th},\ \forall\, t \geq T_0,$$

where $\hat{J}_{th}$ is given by

$$\hat{J}_{th} \triangleq J_{th} + k_1 - \|r(t_b) + C\hat{\phi}(t_b)\|_{\rm RMS}, \tag{48}$$

with k_1 being specified by

$$k_1 \triangleq \sqrt{\frac{\gamma^2}{T_w}\left(\lambda_{\max}(\Theta_0)\delta_1^2 + \Delta^2\right)}. \tag{49}$$

■

Proof. According to (41) and (47), by using the triangle inequality, $\hat{J}\,(t_b|t)$ satisfies

$$\begin{aligned}\hat{J}\,(t_b|t) \leq\ & \|r(t_b)\|_{\rm RMS} + \|e_y(t)\|_{\rm RMS} \\ & + \|\tilde{y}^a(t_b|t)\|_{\rm RMS} - \|r(t_b) + C\hat{\phi}(t_b)\|_{\rm RMS}. \end{aligned}\tag{50}$$

Based on Lemma 1, in the absence of an attack, $\tilde{y}^a(t_b|t) = 0$ and thus, $\hat{J}_{th}$ is chosen as

$$\begin{aligned}\hat{J}_{th} =& \sup_{\tilde{y}^a(t_b|t)=0,\ t\geq t_b} \hat{J}\,(t_b|t) \\ =& \|r(t_b)\|_{\rm RMS} + \sup_{t\geq t_b} \|e_y(t)\|_{\rm RMS} \\ & - \|r(t_b) + C\hat{\phi}(t_b)\|_{\rm RMS}. \end{aligned}\tag{51}$$

Note that when no attack is present, $\|r(t_b)\|_{\rm RMS} \leq J_{th}$. Next, the supremum $\|e_y\|_{\rm RMS}$ will be derived based on the $\mathcal{H}_\infty$ performance of the smoother $\mathcal{S}$. From Theorem 2, one can derive that

$$\|e_y(t)\|^2_{[t_b,t_s]} \leq \gamma^2 \left(e^T(t_b)\Theta e(t_b) + \|d(t)\|^2_{[t_b,t_s]}\right).$$

From (33), (35) and the structure of Θ in (43), we have

$$e^T(t_b)\Theta e(t_b) = e_z^T(t_b)\Theta_0 e_z(t_b) \leq \lambda_{\max}(\Theta_0)\delta_1^2.$$

From $\|d(t)\|_{[t_s,t_b]} \leq \Delta$ and the fact that $\|e_y\|^2_{\rm RMS} \leq \frac{1}{T_w}\|e_y\|^2_{[t_b,t_s]}$, we have $\|e_y\|^2_{\rm RMS} \leq k_1^2$ where k_1 is given by (49). Hence, from (51) and (49), the threshold $\hat{J}_{th}$ in (48) can be obtained. □

Therefore, the above design and analysis can be summarized by the following theorem.

Theorem 3. *(Robustness). Under Assumptions 1-3, when the CPS and the anomaly detector $(\Sigma, \mathcal{D})$ described by (2), (3) and (4) undergo a stealthy integrity attack, the detection decision scheme, characterized by the smoother $\mathcal{S}$ (32) with the optimal parameters (45), residual (40), evaluation function (47) and threshold (48), guarantees that there is no false alarm before the occurrence of the attack, i.e., $\hat{J}\,(t_b|t) \leq \hat{J}_{th}$ for $t_b \leq t < T_0$.* ■

Subsequently, if $\hat{J}(t_b|t) > \hat{J}_{th}$ for some $t \geq T_0 > t_b$, an alarm is triggered and indicates the presence of an attack. The detection time T_d is defined as the first time instant when $\hat{J}(t_b|t) > \hat{J}_{th}$ for a given $t_b < T_0$, i.e.,

$$T_d(t_b) \triangleq \inf\left\{t \geq T_0 \,\middle|\, \hat{J}(t_b|t) > \hat{J}_{th}\right\}. \tag{52}$$

4.4 Detectability Analysis

In this part, the attack detectability analysis is conducted. The detectability analysis constitutes a theoretical result that characterizes quantitively and implicitly the class of stealthy integrity attacks that can be detected by the proposed scheme. The detectability analysis is a theoretical tool that is used correspondingly in the detection of faults, since it provides intuition about the characteristics of detectable faults (see, e.g., Keliris, Polycarpou, and Parisini (2017); Zhang, Polycarpou, and Parisini (2010); Zhang, Jiang, Yan, and Shen (2019); Wu, Jiang, and Lu (2017)).

Theorem 4. *(Detectability) Under Assumptions 1-3, for the CPS and the anomaly detector $(\Sigma, \mathcal{D})$ described in (2), (3) and (4), the attack detection decision scheme, characterized by the smoother $\mathcal{S}$ (32) with the optimal parameters (45), residual (40), evaluation (47) and threshold (48), guarantees that a stealthy integrity attack can be detected at a time $T_d \geq T_0 > t_b$, i.e., $\hat{J}(t_b|T_d) > \hat{J}_{th}$, if the following condition holds:*

$$\|r(t_b|t)\|_{\rm RMS} > J_{th} + 2k_1, \tag{53}$$

where J_{th} is given in (4) and k_1 is defined in (49). ■

Proof. For $\hat{r}\,(t_b|T_d) = r\,(t_b|T_d) - e_y(t)$ given in (41), we have

$$\|\hat{r}(t_b|T_d)\|_{\rm RMS} \geq \|r(t_b|t)\|_{\rm RMS} - \|e_y(t)\|_{\rm RMS}.$$

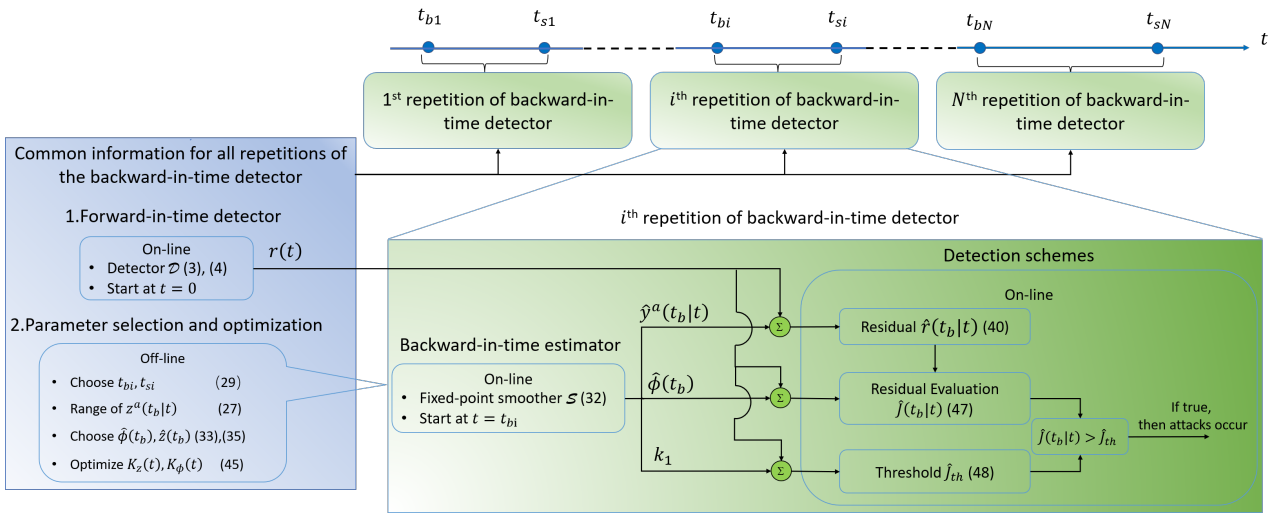

Fig. 4. Diagram of the implementation for the backward-in-time detection scheme.

According to (47) and (48), to detect an attack at the time instant T_d, the following inequality should hold:

$$\|\hat{r}\,(t_b|T_d)\,\|_{\text{RMS}} > J_{th} + k_1.$$

Then, a sufficient condition can be obtained as follows:

$$\|r(t_b|t)\|_{\text{RMS}} > J_{th} + k_1 + \|e_y(t)\|_{\text{RMS}}.$$

Hence, from $\|e_y\|_{\text{RMS}} \leq k_1$, the sufficient condition (53) follows. □

Remark 11. Compared with the inequality (26) in the result (b) of Theorem 1, (53) has the additional term $2k_1$, which is the result of using the fixed-point smoother to estimate $z^a(t_b|t)$. Note that k_1 defined in (49) is the upper bound of $\|e_y\|_{\text{RMS}}$. Based on $\mathcal{H}_\infty$ property of the estimation error system (39), k_1 in the right hand side of (53) is fixed in the presence of any disturbance satisfying $\|d\|_{[t_b,t_s]} \leq \Delta$. It should also be noted that $r(t_b|t)$ in (53) can be increased to any required value in the presence of the attack via selecting a finite t_b (see result (b) in Theorem 1). Therefore, the detectability of the proposed attack detection methodology can be as high as possible by making $r(t_b|t)$ sufficiently large via the selection of a finite t_b. ∇

Fig. 4 illustrates the implementation of the backward-in-time attack detection scheme developed in this paper. A common information block that includes a forward-in-time detector $\mathcal{D}$ and the parameter selection and optimization scheme, is shared among the N repetitions of the backward-in-time detectors (blue box). The anomaly detector $\mathcal{D}$ is activated once at $t = 0$ and is always in parallel with all repetitions. Each repetition includes a smoother (32), which indicates that the smoother (32) is activated N times and the differential Riccati equation (44) is solved N times. The common parameter selection and optimization block is done off-line and provides every repetition of the backward-in-time detector with the necessary information before it is initiated. The ith backward-in-time detector includes a backward-in-time estimator block and a detection scheme block that run on-line during the time interval $[t_{b,i}, t_{s,i}]$ in real time, which provides the occurrence information of attacks that potentially occur in the time interval $[t_{b,i}, t_{s,i}]$.

5 Case Study

In this section, the longitudinal navigation mathematical model of an air breathing hypersonic vehicle is considered. Such a model is a simple numerical example for illustration purposes. Unmanned hypersonic vehicles are vulnerable to such as GPS spoofing attacks which can affect not only the control input but also the output measurements of the navigation system. Based on the control-design model in Fiorentini et al. (2009) and by considering altitude, angle of attack and pitch rate as the states, and sinusoidal and cosine of flight-path angle as the control inputs, the longitudinal navigation dynamics can be written in the form Σ where $x = [x_1, x_2, x_3]^T$ with x_1, x_2 and x_3 representing the altitude, angle of attack and pitch rate, respectively. The system matrices are given as follows:

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0.7586 & 10^6 \\ 0 & 2.6489 & -1.6197 \end{bmatrix} \times 10^{-6},\ D_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix},$$
$$B = \begin{bmatrix} 1.5 & 0 \\ 0 & 6.5333 \\ 0 & 0 \end{bmatrix} \times 10^{-4},\ C = \begin{bmatrix} 0.1 & 0 & 0 \\ 0.1 & 1000 & 0 \end{bmatrix}, D_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}.$$

It can be verified that the pair (A, C) is observable. In addition, the nonlinear function is $g(t, x) = [0, g_2(t, x), g_3(t, x)]^T$ where $g_2(t, x) = 3.6412\times10^{-9}\sin(x_2)$ and $g_3(t, x) = -3.0475 \times 10^{-4}x_2^2 - 4.8088 \times 10^{-5}x_2^2x_3 + 2.1334 \times 10^{-6}x_2x_3$.

Suppose that the available resources to the attacker is

$\Gamma_u = \Gamma_y = I_2$. Then, we can calculate that $x_1(t) \in \text{Im}[1,0,0]^T$ satisfies (7a) and (7b) and thus, Assumption 1 is satisfied. Moreover, the control law u_c is given by $u_c(t,\tilde{y},y_{\text{ref}}) = K_1\int_0^t([1,0]\tilde{y} - y_{\text{ref}})dt + K_2\tilde{y}$, where the reference signal is $y_{\text{ref}} = 1.1\times 10^5$ ft and

$$K_1 = \begin{bmatrix} 0.67\times 10^{-4} \\ 0 \\ 0 \end{bmatrix},\ K_2 = \begin{bmatrix} -0.11739 & 0 \\ 0 & -2.6955\times 10^6 \\ 0 & -517.2942 \end{bmatrix}.$$

It can be verified that the control law u_c can asymptotically stabilize the navigation system in the non-attack case and $y_{\text{ref}} \equiv 0$ for $t \geq 0$, and hence, Assumption 2 is satisfied. In addition, in the region $\mathcal{X} = [0, 135000] \times [-\pi/3, \pi/3] \times [-\pi/3, \pi/3]$, we have $\|g(t,x) - g(t,\hat{z})\| \leq 0.0183\|x - \hat{z}\|$ and thus, the Lipschitz constant l in Assumption 3 is $l = 0.0183$ and Assumption 3 holds.

For the simulation purpose, the initial condition of the state x is given by $[1.0\times 10^6$ ft$, 0.1$ rad$, 0.2$ rad$]^T$, the disturbance is given by $d(t) = [0.003\cos(3t + 0.2), 180 + 20\sin(40t), 0.063\sin(10t + 0.2)]^T$. Thus, we have $\|d(t)\|_{[0,10]} \leq 2000$ and Δ is 2000. The anomaly detector $\mathcal{D}$ is designed based on Ding (2013). The residual is designed to satisfy the optimal $\mathcal{H}_\infty$ performance, i.e., $\|r(t)\|_{[0,10]} \leq 11\|d(t)\|_{[0,10]}$, and thus the threshold is calculated as $J_{th} = 11\Delta/\sqrt{T_w} = 3.113\times 10^4$ where $T_w = 0.5$. The attack signal $a(t)$ used in this simulation is generated by the attack model in *Remark 4* where z_0 and F_a are chosen as

$$z_0 = \begin{bmatrix} 2000 \\ 0 \\ 0 \end{bmatrix},\ F_a = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -0.1 & 0 & 0 \\ -0.1 & 0 & 0 \end{bmatrix}.$$

Such an attack is initiated at $T_0 = 6$ s. The attack signal is shown in Fig. 5, and $y_{\text{ref}}(t)$, $\tilde{y}(t)$ and $\tilde{y}^n(t)$ are presented in Fig. 6. The detection results using the anomaly detector $\mathcal{D}$ is shown in Fig. 7.

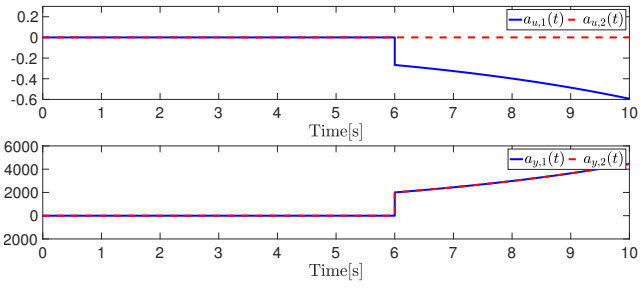

Fig. 5. Time responses of the attack signal $a(t)$.

Fig. 6 shows that the control law u_c can drive the output $\tilde{y}_1$ to the reference y_{ref} asymptotically, which indicates that the control law u_c achieves its objective. It can also be observed from Fig. 6 that the altitude measurement $\tilde{y}_1$ changes slightly, the increment $\tilde{y}_1^a$ is small enough and converges to zero asymptotically. Moreover, the increment $\tilde{y}_2^a$ is zero identically. As shown in Fig. 7, such

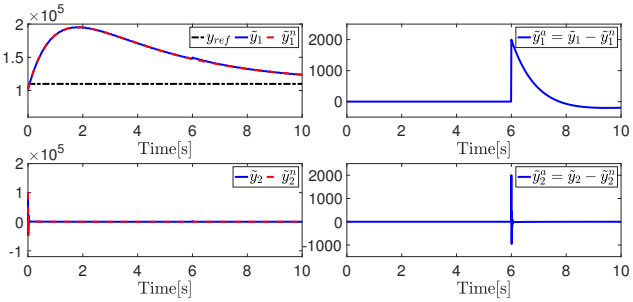

Fig. 6. Time responses of the sensor measurements $\tilde{y}$ and $\tilde{y}^n$ received from $\mathcal{N}_s$ in the attack case and healthy case respectively, the reference signal y_{ref}, and the change $\tilde{y}^a$.

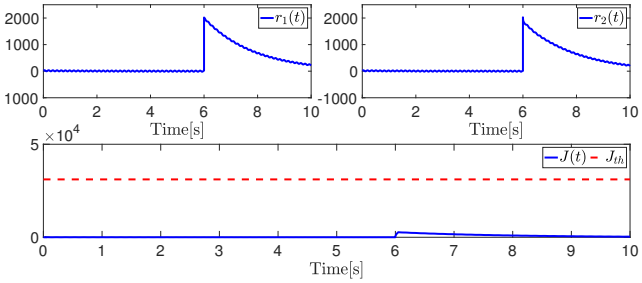

Fig. 7. Time responses of the residual $r(t)$, the evaluation function $J(t)$ and the threshold J_{th} of the anomaly detector $\mathcal{D}$.

changes can not be detected by the anomaly detector $\mathcal{D}$ since $J(t) < J_{th}$ for $t > T_0$. Therefore, such an attack satisfies Definition 1 and remains stealthy with respect to the detector $\mathcal{D}$.

The time t_b is chosen as $t_b = 2$ s. At the time t_b, $r(t_b) = [-13.95, -14.28]^T$. Based on the designed u_c, the related comparison function $\beta(\|z_0\|, t - T_0)$ in (17) is

$$\beta(\|z_0\|, t - T_0) = \exp(-0.4(t - T_0))\|z_0\|.$$

The time $t_s = 10$ s, then, $\rho_{t_s}(x)$ and $\rho_{t_s}^{-1}(x)$ based on the function β are respectively derived by

$$\rho_{t_s}(x) = \exp(-3.3)x,\ \rho_{t_s}^{-1}(x) = \exp(3.3)x.$$

With respect to the given anomaly detector $\mathcal{D}$, it can be determined through simulation that if $\|z_0\| \leq 30000$, then the integrity attacks generated by the attack model in *Remark 4* are stealthy. Thus, the defender knows that the attacker has to select a z_0 satisfying $\|z_0\| < 30000$ in (8), to maintain the stealthiness of the integrity attacks with respect to the anomaly detector $\mathcal{D}$. Hence, in this simulation, we select $\delta_0 = 30000$. Then, $\rho_{t_s}^{-1}(\delta_0) = \exp(3.3)\delta_0 = 8.1338\times 10^5$. Thus, based on Theorem 1, we have

$$7.1720\times 10^5 < \|z^a(t_b|t)\| \leq 8.1338\times 10^5.$$

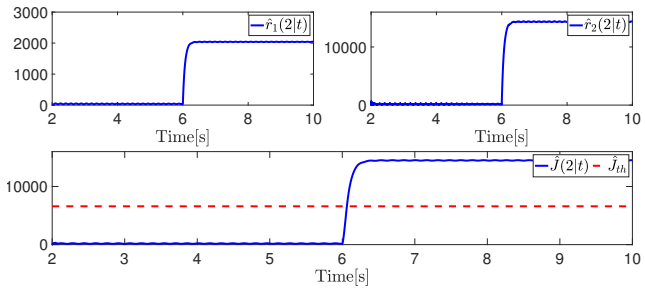

Fig. 8. Time responses of the residual $\hat{r}(2|t)$, the evaluation function $\hat{J}(2|t)$ and the threshold $\hat{J}_{th}$.

In the following, we proceed with the design of the smoother $\mathcal{S}$. The initial condition $\hat{z}(t_b)$ is chosen as $\hat{z}(t_b) = [7.8294 \times 10^5, 0.3913, 0.3358]^T$ and $\hat{\phi}(t_b)$ is chosen as $\hat{\phi}(t_b) = [8.1000 \times 10^5, 0, 0]^T$. The $\mathcal{H}_\infty$ performance index is given by $\gamma = 15$ and the Θ_0 is selected as $\Theta_0 = 0.605 I_3$. By solving the differential Riccati equation (44), the solution $Q(t)$ is obtained. Thus, $K_z(t)$ and $K_\phi(t)$ are determined based on the solution $Q(t)$. Thus, the optimal fixed-point smoother is determined as well and $\tilde{y}^a(z|t)$ is estimated. Therefore, the residual is

$$\hat{r}(2|t) = r(2) + \hat{\tilde{y}}^a(2|t).$$

Moreover, based on (49), $k_1 = 9.7540 \times 10^4$ and hence based on (48), the threshold $\hat{J}_{th}$ is $\hat{J}_{th} = 6.6268 \times 10^3$.

The detection results using the above determined backward-in-time detector are shown in Fig. 8. It can be seen that the stealthy integrity attack is detected at about $T_d = 6.2$s, since the residual evaluation function $\hat{J}(2|T_d)$ exceeds the threshold $\hat{J}_{th}$. In addition, the residuals $\hat{r}_1(2|t)$ and $\hat{r}_2(2|t)$ shown in Fig. 8, regardless of the fluctuations due to disturbances, retain their values as constant vectors, which verifies the result (a) in Lemma 1.

6 Conclusions

In this paper, a stealthy integrity attack detection methodology has been proposed for a class of nonlinear CPS. An equivalent increment of the system at a time prior to the attack occurrence time has been defined and its effectiveness to detect stealthy integrity attacks has been investigated. A backward-in-time detector based on an $\mathcal{H}_\infty$ fixed-point smoother has been proposed as the tool to estimate the unknown equivalent increment. Based on the aforementioned findings, the detection scheme has been designed and rigorously investigated by conducting a detectability analysis. Finally, a simulation case study has presented to show the effectiveness of the developed detection methodology. Future research efforts will be devoted to distinguish between fault anomalies and stealthy integrity attacks.

Appendix: Proof of Theorem 2

Proof. Based on Assumption 3, $\|\bar{g}(t,z,\hat{z})\| \le l\|e_z(t)\|$, and it then follows from $e^T(t) = [e_z^T(t), e_\phi^T(t)]$ that

$$\bar{g}^T(t,z,\hat{z})\bar{g}(t,z,\hat{z}) \le l^2 e^T(t) E_{n_x} e(t). \tag{A.1}$$

From (42), the $\mathcal{H}_\infty$ performance can be guaranteed if $J(\gamma) \le 0$ in the worst case where

$$J(\gamma) = \int_{t_b}^{t_s} \gamma^{-2} e_y^T(t) e_y(t) - d(t)^T d(t)\, dt - e_0^T \Theta e_0.$$

In the sequel, the task is to prove that with the $K(t)$ given in Theorem 2, $J(\gamma) \le 0$ in the worst case.

From $(A.1)$ and along the dynamics of the estimation error system (39), we can obtain

$$\begin{aligned}
J(\gamma) = &\int_{t_b}^{t_s} [\frac{d\, e^T P e}{dt} + \gamma^{-2} e_y^T e_y - d^T d]\, dt + e^T P(t_b) e \\
&- e^T P(t_s) e - e_0^T \Theta e_0 \\
\le &\int_{t_b}^{t_s} [e^T(\dot{P} + (\bar{A} + K\bar{C}_1)^T P + P(\bar{A} + K\bar{C}_1))e \\
&+ \mu^2 e^T P E_{n_x} P e + e^T(\mu^{-2} l^2 E_{n_x} + \gamma^{-2} \bar{C}_2^T \bar{C}_2) e \\
&+ e^T P(\bar{D}_1 + KD_2)(\bar{D}_1 + KD_2)^T P e \\
&- (d - e^T P(\bar{D}_1 + KD_2))^T \cdot \\
&\qquad (d - e^T P(\bar{D}_1 + KD_2))]\, dt \\
&+ e^T P(t_b) e - e^T P(t_s) e - e_0^T \Theta e_0,
\end{aligned}$$

where the following inequality is used:

$$\begin{aligned}
2e^T P \bar{g} &\le \mu^2 e^T P [I_{n_x}, 0]^T [I_{n_x}, 0] P e + \mu^{-2} \bar{g}^T \bar{g} \\
&= \mu^2 e^T P E_{n_x} P e + \mu^{-2} e^T E_{n_x} e.
\end{aligned}$$

We now choose a $P(t)$ satisfying the following differential Riccati equation on the time interval $[t_b, t_s]$:

$$\begin{aligned}
-\dot{P} = &\left(\bar{A} + K\bar{C}_1\right)^T P + P\left(\bar{A} + K\bar{C}_1\right) \\
&+ P\left(\mu^2 E_{n_x} + (\bar{D}_1 + KD_2)(\bar{D}_1 + KD_2)^T\right) P \\
&+ \mu^{-2} l^2 E_{n_x} + \gamma^{-2} \bar{C}_2^T \bar{C}_2, \qquad (A.2) \\
P(t_b) = &\Theta + \Delta\Theta, \qquad (A.3)
\end{aligned}$$

where $\Delta\Theta$ is introduced to avoid the singularity of $P(t_b)$, which is given as follows:

$$\Delta\Theta = \frac{1}{\varepsilon} \begin{bmatrix} I & -I \\ -I & I \end{bmatrix}, \ \varepsilon \to 0.$$

It then can be verified that in the worst case $d = e^T P(\bar{D}_1 + KD_2)$, $J(\gamma) \le 0$. Hence, the inequality (42) follows.

In the following, the procedure to derive (44) and (46) from ($A.2$) and ($A.3$) is presented. Let $Q(t) = P^{-1}(t)$ for $t \in [t_b, t_s]$. Then,

$$Q(t_b) = \lim_{\varepsilon \to 0} (\Theta + \Delta\Theta)^{-1} = \begin{bmatrix} \Theta_0^{-1} & \Theta_0^{-1} \\ \Theta_0^{-1} & \Theta_0^{-1} \end{bmatrix}. \quad (A.4)$$

Moreover, it follows from ($A.2$) that

$$\begin{aligned} \dot{Q} = {} & Q\bar{A}^T + \bar{A}Q + Q\left(\mu^{-2} l^2 E_{n_x} + \gamma^{-2} \bar{C}_2^T \bar{C}_2\right) Q \\ & + \mu^2 E_{n_x} + \bar{D}_1 \bar{D}_1^T \\ & + \left(K^T - R^{-1}\left(\bar{C}_1 Q + D_2 \bar{D}_1^T\right)\right)^T R \cdot \\ & \qquad \left(K^T - R^{-1}\left(\bar{C}_1 Q + D_2 \bar{D}_1^T\right)\right) \\ & - \left(\bar{C}_1 Q + D_2 \bar{D}_1^T\right)^T R^{-1} \left(\bar{C}_1 Q + D_2 \bar{D}_1^T\right). \quad (A.5) \end{aligned}$$

Then, by substituting $K(t)$ in (46) into ($A.5$), (44) can be obtained.

Hence, Theorem 2 is proved. □

Acknowledgements

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

References

An, L. and Yang, G. (2017). Data-driven coordinated attack policy design based on adaptive L_2-gain optimal theory. *IEEE Transactions on Automatic Control*, 63(6), 1850–1857.

Angeli, D. (2002). A Lyapunov approach to incremental stability properties. *IEEE Transactions on Automatic Control*, 47(3), 410–421.

Barboni, A., Rezaee, H., Boem, F., and Parisini, T. (2020). Detection of covert cyber-attacks in interconnected systems: A distributed model-based approach. *IEEE Transactions on Automatic Control*, 65(9), 3728–3741.

Blanke, M., Kinnaert, M., Lunze, J., Staroswiecki, M., and Schröder, J. (2006). *Diagnosis and fault-tolerant control*. Springer-Verlag Berlin Heidelberg.

Cardenas, A., Amin, S., and Sastry, S. (2008). Secure control: Towards survivable cyber-physical systems. In *28th International Conference on Distributed Computing Systems Workshops*, 495–500. IEEE.

Chen, J. and Patton, R. (1999). *Robust model-based fault diagnosis for dynamic systems*. Springer Science & Business Media.

Chen, Y., Kar, S., and Moura, J. (2016). Dynamic attack detection in cyber-physical systems with side initial state information. *IEEE Transactions on Automatic Control*, 62(9), 4618–4624.

Dibaji, S., Pirani, M., Flamholz, D., Annaswamy, A., Johansson, K., and Chakrabortty, A. (2019). A systems and control perspective of CPS security. *Annual Reviews in Control*, 47, 394–411.

Ding, S. (2013). *Model-based fault diagnosis techniques: design schemes, algorithms, and tools*. Springer-Verlag London.

Einicke, G. (2019). *Smoothing, filtering and prediction: Estimating the past, present and future*. InTech.

Ferrari, R. and Teixeira, A. (2017). Detection and isolation of replay attacks through sensor watermarking. *IFAC-PapersOnLine*, 50(1), 7363–7368.

Ferrari, R. and Teixeira, A. (2021). A switching multiplicative watermarking scheme for detection of stealthy cyber-attacks. *IEEE Transactions on Automatic Control*, 66(6), 2558–2573.

Fiorentini, L., Serrani, A., Bolender, M., and Doman, D. (2009). Nonlinear robust adaptive control of flexible air-breathing hypersonic vehicles. *Journal of Guidance, Control, and Dynamics*, 32(2), 402–417.

Griffioen, P., Weerakkody, S., and Sinopoli, B. (2021). A moving target defense for securing cyber-physical systems. *IEEE Transactions on Automatic Control*, 66(5), 2016–2031.

Hoehn, A. and Zhang, P. (2016). Detection of covert attacks and zero dynamics attacks in cyber-physical systems. In *American Control Conference*, 302–307. IEEE.

Isidori, A. (2013). *Nonlinear control systems*. Springer Science & Business Media.

Keliris, C., Polycarpou, M., and Parisini, T. (2017). An integrated learning and filtering approach for fault diagnosis of a class of nonlinear dynamical systems. *IEEE transactions on Neural Networks and Learning Systems*, 28(4), 988–1004.

Khalil, H. (2001). *Nonlinear Systems*. Pearson.

Meditch, J. (1967). On optimal fixed point linear smoothing. *International Journal of Control*, 6(2), 189–199.

Milošević, J., Teixeira, A., Johansson, K., and Sandberg, H. (2020). Actuator security indices based on perfect undetectability: Computation, robustness, and sensor placement. *IEEE Transactions on Automatic Control*, 65(9), 3816–3831.

Mo, Y., Chabukswar, R., and Sinopoli, B. (2013). Detecting integrity attacks on SCADA systems. *IEEE Transactions on Control Systems Technology*, 22(4), 1396–1407.

Mo, Y. and Sinopoli, B. (2009). Secure control against replay attacks. In *47th annual Allerton conference on communication, control, and computing*, 911–918. IEEE.

Na, G. and Eun, Y. (2018). A multiplicative coordinated stealthy attack and its detection for cyber physical systems. In *2018 IEEE Conference on Control Technology and Applications (CCTA)*, 1698–1703. IEEE.

Pasqualetti, F., Dörfler, F., and Bullo, F. (2013). Attack detection and identification in cyber-physical systems. *IEEE Transactions on Automatic Control*, 58(11), 2715–2729.

Pasqualetti, F., Dorfler, F., and Bullo, F. (2015). Control-theoretic methods for cyber physical security: Geometric principles for optimal cross-layer resilient control systems. *IEEE Control Systems Magazine*, 35(1), 110–127.

Romagnoli, R., Weerakkody, S., and Sinopoli, B. (2019). A model inversion based watermark for replay attack detection with output tracking. In *American Control Conference*, 384–390. IEEE.

Sánchez, H., Rotondo, D., Escobet, T., Puig, V., and Quevedo, J. (2019). Bibliographical review on cyber attacks from a control oriented perspective. *Annual Reviews in Control*, 48, 103–128.

Shaked, U. and Theodor, Y. (1992). H_∞-optimal estimation: a tutorial. In *31st IEEE Conference on Decision and Control*, 2278–2286. IEEE.

Simon, D. (2006). *Optimal state estimation: Kalman, H_∞, and nonlinear approaches*. John Wiley & Sons.

Smith, R. (2011). A decoupled feedback structure for covertly appropriating networked control systems. *IFAC Proceedings Volumes*, 44(1), 90–95.

Smith, R. (2015). Covert misappropriation of networked control systems: Presenting a feedback structure. *IEEE Control Systems Magazine*, 35(1), 82–92.

Teixeira, A., Shames, I., Sandberg, H., and Johansson, K. (2012). Revealing stealthy attacks in control systems. In *50th Annual Allerton Conference on Communication, Control, and Computing*, 1806–1813. IEEE.

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

Teixeira, A., Sou, K., Sandberg, H., and Johansson, K. (2015b). Secure control systems: A quantitative risk management approach. *IEEE Control Systems Magazine*, 35(1), 24–45.

Theodor, Y. and Shaked, U. (1994). Game theory approach to H_∞-optimal discrete-time fixed-point and fixed-lag smoothing. *IEEE Transactions on Automatic Control*, 39(9), 1944–1948.

Trentelman, H., Stoorvogel, A., and Hautus, M. (2012). *Control theory for linear systems*. Springer-Verlag London.

Weerakkody, S., Ozel, O., Griffioen, P., and Sinopoli, B. (2017). Active detection for exposing intelligent attacks in control systems. In *2017 IEEE Conference on Control Technology and Applications (CCTA)*, 1306–1312. IEEE.

Weerakkody, S. and Sinopoli, B. (2015). Detecting integrity attacks on control systems using a moving target approach. In *54th IEEE Conference on Decision and Control*, 5820–5826. IEEE.

Wu, Y., Jiang, B., and Lu, N. (2017). A descriptor system approach for estimation of incipient faults with application to high-speed railway traction devices. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 49(10), 2108–2118.

Ye, D. and Zhang, T. (2019). Summation detector for false data-injection attack in cyber-physical systems. *IEEE Transactions on Cybernetics*, 50(6), 2338–2345.

Zemouche, A., Boutayeb, M., and Bara, G. (2005). Observer design for nonlinear systems: An approach based on the differential mean value theorem. In *44th IEEE Conference on Decision and Control*, 6353–6358. IEEE.

Zhang, K., Jiang, B., Yan, X., and Shen, J. (2019). Interval sliding mode observer based incipient sensor fault detection with application to a traction device in China railway high-speed. *IEEE Transactions on Vehicular Technology*, 68(3), 2585–2597.

Zhang, K., Polycarpou, M.M., and Parisini, T. (2020). Enhanced anomaly detector for nonlinear cyber-physical systems against stealthy integrity attacks. *IFAC-PapersOnLine*, 53(2), 13682–13687.

Zhang, T. and Ye, D. (2020). False data injection attacks with complete stealthiness in cyber–physical systems: A self-generated approach. *Automatica*, 120, 109117.

Zhang, X., Polycarpou, M., and Parisini, T. (2010). Fault diagnosis of a class of nonlinear uncertain systems with Lipschitz nonlinearities using adaptive estimation. *Automatica*, 46(2), 290–299.