Stealthy Integrity Attacks for a Class of Nonlinear Cyber-Physical Systems

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

***Abstract*—This paper proposes a stealthy integrity attack generation methodology for a class of nonlinear cyber-physical systems. Geometric control theory and stability theory of incremental systems are used to design an attack generation scheme with stealthiness properties. An attack model is proposed as a closed-loop dynamical system with an arbitrary input signal. This model is developed based on a controlled invariant subspace that results from geometric control theory and is decoupled with the system outputs and the nonlinear function. The presence of the arbitrary signal in the attack model provides an additional degree of freedom and constitutes a novel component compared with existing results. The stealthiness of the attack model is rigorously investigated based on the incremental stability of the closed-loop control system, and the incremental input-to-state stability of the anomaly detector. As a result, a sufficient condition in terms of the initial condition of the attack model is derived to guarantee stealthiness. Finally, a case study is presented to illustrate the effectiveness of the developed attack generation scheme.**

***Index Terms*—Integrity attacks, attack model, nonlinear cyber-physical systems.**

I. Introduction

Cyber-physical systems (CPS) have attracted significant attention recently as a result of their wide applications. CPS integrate computing, communication and control [1], which are highly vulnerable to malicious cyber attacks [2]. Therefore, motivated by security and safety specifications, investigating cyber attacks becomes critical for defense technologies.

Attacks are usually generated by rational adversary models and empowered with intelligence and intent. Associated researches in attack generation pave the way to deeply understand attack behaviors, thereby helping defenders against malicious attacks. In the past decade, the survey papers [3]–[7] and the references therein, have detailed the research directions for the attack generation from a control perspective. Generally, cyber attacks are classified into two categories: denial of service (DoS) attacks and integrity (or deception) attacks. DoS attacks are performed through compromising the *availability* of the data transmission networks [1]. *Integrity attacks*, including replay attacks [8], covert attacks [9], [10], zero-dynamics attacks [3], inject false data to communication networks for achieving attack targets.

Integrity attacks are among the most researched attacks by the control community. The model-knowledge-based attacks, such as covert attacks and zero-dynamics attacks, attract more attention as a result that the related theories are not yet well established. Zero-dynamics attacks for strictly proper linear systems are designed by using a controlled invariant subspace decoupled from the system output in [11]. A weakly unobservable subspace is exploited to design zero-dynamics attacks for proper linear systems in [12]. In [13], robust stealthy zero-dynamics attacks are generated for uncertain linear systems by replacing the real zero dynamics with some nominal zero dynamics. For linear systems with a topology-switching attack defense strategy, [14] proposes a type of zero-dynamics attacks that can bypass the defense strategy without being detected. However, the aforementioned works mainly focus on linear systems. To the authors' best knowledge, [15] is the only attack generation work for nonlinear systems, where the nonlinear system is limited to the Byrnes-Isidori form. The objective of this paper is to propose a stealthy integrity attack generation methodology for a class of nonlinear systems.

Specifically, by extending the authors' previous work [16], a stealthy integrity attack generation approach is proposed for a class of CPS with nonlinear physical plant. The attacks are formulated in the context of nonlinear CPS such that the generated attacks are stealthy with respect to typical anomaly detectors. The generation approach is achieved by using geometric control theory [17], and stability theory of incremental systems [18]. In particular, the attack generation model is proposed as a closed-loop system with an arbitrary input signal. Such a model is based on a controlled invariant subspace decoupled with the system outputs and the nonlinear function. The arbitrary input signal constitutes a novel contribution by providing an alternative excitation, which is particularly useful when the initial condition of the model is close or equal to zero. The stealthiness of the generated attack is rigorously investigated based on the incremental stability of the closed-loop system, and the incremental input-to-state stability of the anomaly detector. A sufficient condition on the initial value of the attack model is derived, allowing the generated attacks to remain undetected by typical anomaly detectors.

This note is organized as follows. In Section II, the problem formulation is given. In Section III, the stealthy integrity attack

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

Kangkang Zhang, Christodoulos Keliris and Marios 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 kzhang02@ucy.ac.cy, keliris.chris@gmail.com, mpolycar@ucy.ac.cy

T. Parisini is with the Dept. of Electrical and Electronic Engineering at Imperial College London, UK, with the KIOS Research and Innovation Centre of Excellence, University of Cyprus, and also with the Dept. of Engineering and Architecture at the University of Trieste, Italy t.parisini@gmail.com

model is presented in detail and in Section IV, a case study is illustrated. Finally, some conclusions are drawn in Section V.

Notation: For a set S, $|S|$ represents the number of the elements in S. Considering a vector signal $x(t) : \mathbb{R}_{\geq 0} \to \mathbb{R}^n$, $x(t) \equiv 0$ 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 $t \in [t_1, t_2]$. For a linear map $A : \mathcal{X} \longrightarrow \mathcal{Y}$, we define $\ker A \triangleq \{x \in \mathcal{X} \mid Ax = 0\}$ and $\text{Im}A \triangleq \{Ax \mid x \in \mathcal{X}\}$.

II. Problem Formulation

A general structure of CPS subject to integrity type of cyber attacks is depicted in Fig. 1, which consists of a physical plant $\mathcal{P}$, a feedback controller $\mathcal{C}$, an anomaly detector $\mathcal{D}$, actuator and sensor communication networks $\mathcal{N}_a$ and $\mathcal{N}_s$ respectively. The attack generation block compromises the communication networks $\mathcal{N}_a$ and $\mathcal{N}_s$ by injecting additive false data a_u and a_y respectively designed by the adversary based on the specific attack model. Suppose that the CPS has n_u actuator communication channels and n_y sensor communication channels. 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. Then, $a_u \in \mathbb{R}^{|K_u|}$ and $a_y \in \mathbb{R}^{|K_y|}$. For each $i \in \{1, \cdots, |K_u|\}$ ($j \in \{1, \cdots, |K_y|\}$), $a_{u,i}(t) \equiv 0$ ($a_{y,j}(t) \equiv 0$) for $t \geq 0$ if there is no attack occurring on the i-th $\mathcal{N}_a$ (j-th $\mathcal{N}_s$) channel. In the sequel, the combined attack vector is denoted as $a(t) \triangleq [a_u^T(t), a_y^T(t)]^T \in \mathbb{R}^{|K_u|+|K_y|}$.

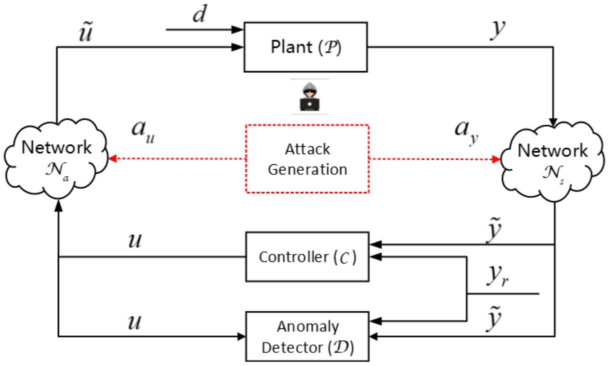

Fig. 1. General architecture of CPSs under potential integrity cyber attacks.

The closed-loop CPS consisting of $\mathcal{C}$, $\mathcal{P}$, $\mathcal{N}_a$ and $\mathcal{N}_s$ is denoted by $\mathcal{W}$ throughout the paper, and in the attack case, $\mathcal{W}$ is described by

$$\mathcal{W} : \begin{cases} \dot{x}(t) = Ax(t) + g(t, x) + B\tilde{u}(t) + D_1 d(t), & \text{(1a)} \\ \tilde{u}(t) = u(t, \tilde{y}, y_{\text{ref}}) + \Gamma_u a_u(t), & \text{(1b)} \\ y(t) = Cx(t) + D_2 d(t), & \text{(1c)} \\ \tilde{y}(t) = y(t) + \Gamma_y a_y(t), & \text{(1d)} \end{cases}$$

where $x \in \mathbb{R}^{n_x}$ is the state vector, $\tilde{u}$, $u \in \mathbb{R}^{n_u}$ are the control data received by the actuator and computed by the controller $\mathcal{C}$ respectively, $\tilde{y}$, $y \in \mathbb{R}^{n_y}$ denote the sensor measurements received by the controller $\mathcal{C}$ and the outputs of the physical plant $\mathcal{P}$ respectively. The signal $y_{\text{ref}} \in \mathbb{R}^{n_{y_{\text{ref}}}}$ is the reference signal, and $d \in \mathbb{R}^{n_d}$ represents the lumped disturbances and noise. 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 Γ_u and Γ_y are related to the disruption resources K_u and K_y respectively, and also indicate the actuator and sensor communication channels that can be affected by the adversary. The function $g : \mathbb{R}_{\geq 0} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$ represents the nonlinearity, which is piecewise continuous in t and continuous differentiable in x, and also satisfies $g(t, 0) = 0$. The function $u : \mathbb{R}_{\geq 0} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_y} \to \mathbb{R}^m$ is the nonlinear output feedback control law such that in the nominal case ($d(t) \equiv 0$ and $a(t) \equiv 0$ for $t \geq 0$), the output y asymptotically tracks y_{ref}. Moreover, u is piecewise continuous in t and $u(t, 0, 0) = 0$.

The anomaly detector $\mathcal{D}$ in Fig. 1 takes the form of a commonly used detector such as one of the model-based detectors in [19]. Specifically, $\mathcal{D}$ has the following form:

$$\mathcal{D} : \begin{cases} \dot{x}_r(t) = A_r x_r(t) + g(t, x_r) + Bu(t) + K_r \tilde{y}(t), & \text{(2a)} \\ r(t) = Cx_r(t) - \tilde{y}(t), & \text{(2b)} \\ J(t) = \|r(t)\|_{\text{RMS}}, & \text{(2c)} \\ J_{th} = \bar{J}_{th}\left(d(t)\right), & \text{(2d)} \end{cases}$$

where $x_r \in \mathbb{R}^{n_x}$ is the state of the detector and $r \in \mathbb{R}^{n_y}$ is the *residual* vector used for the anomaly detection. Moreover, $J(t)$ is the evaluation function that is chosen as the root-mean-square (RMS[1]) value of $r(t)$. The $J_{th} \in \mathbb{R}$ is the detection *threshold*, and is determined by the function $\bar{J}_{th} : \mathbb{R}^{n_d} \to \mathbb{R}_{\geq 0}$ that is a scalar function of the disturbance. The gain matrix $K_r \in \mathbb{R}^{n_x \times n_y}$ is chosen such that $A_r \triangleq A - K_r C$ is Hurwitz and the following assumption is also satisfied.

Assumption 1. The differential equation (2a) in $\mathcal{D}$ is incremental input-to-state stable (i-ISS) with respect to u and $\tilde{y}$, i.e., there exists a class $\mathcal{KL}$ function β_1 and class $\mathcal{K}_\infty$ functions ρ_1 and ρ_2 such that for any two initial conditions $x_{r,0}^1$ and $x_{r,0}^2$, two input pairs $(u_1, \tilde{y}_1)$ and $(u_2, \tilde{y}_2)$ (inputs of (2a)), the following inequality holds:

$$\begin{aligned} \|x_r^1(t) - x_r^2(t)\| &\leq \beta_1(\|x_{r,0}^1 - x_{r,0}^2\|, t) \\ &\quad + \rho_1(\|u_1(t) - u_2(t)\|) + \rho_2(\|\tilde{y}_1(t) - \tilde{y}_2(t)\|),\ \forall\, t \geq 0, \end{aligned}$$

where x_r^1 and x_r^2 are solutions of (2a) resulting from the excitation triples $(x_{r,0}^1, u_1, \tilde{y}_1)$ and $(x_{r,0}^2, u_2, \tilde{y}_2)$ respectively.

Remark 1. The i-ISS is used to characterize the boundedness of the increments of the system states due to the increments of the system inputs and initial conditions [18]. Given the anomaly detector $\mathcal{D}$ in (2), the increment of the residual $r(t)$, resulting from the increments of $u(t)$ and $\tilde{y}(t)$ due to attacks, results in the detection of an attack. Hence, the i-ISS concept is introduced to characterize the boundedness of the increment of $r(t)$ in the presence of attacks resulting from the input increments of $u(t)$ and $\tilde{y}(t)$ in the anomaly detector $\mathcal{D}$. ∇

Remark 2. The i-ISS is equivalent to the existence of an i-ISS-Lyapunov function (see Theorem 2 in [18]). From the i-ISS-Lyapunov function in [18], we can deduce that if the estimation error $x - x_r$ asymptotically converges to zero in the nominal

[1] The RMS is defined as $\|r(t)\|_{\text{RMS}} \triangleq \left(\frac{1}{T_w}\int_{t-T_w}^{t} r^T(\tau)r(\tau)d\tau\right)^{\frac{1}{2}}$ where $T_w > 0$ is the length of the time window

case ($d(t) \equiv 0$ and $a(t) \equiv 0$ for $t \geq 0$), then Assumption 1 holds. Consider a special case that $g(t,x)$ is locally Lipschitz in a compact set $\mathcal{X} \subset \mathbb{R}^{n_x}$, i.e., $\|g(t,x) - g(t,x_r)\| \leq l\|x - x_r\|$. A sufficient condition that the error $x - x_r$ converges asymptotically to zero is given in [20], which states that there exists a matrix $P = P^T > 0$ such that $A_r^T P + PA_r + l^2 PP + I + \varepsilon I = 0$, where ε can be any positive scalar. ∇

The typical anomaly detector $\mathcal{D}$ is introduced for the purpose of defining stealthy attacks. The occurrence of an attack is ascertained by $\mathcal{D}$ if $J(t) > J_{th}$ for some time instant. Otherwise, the system is considered to be in normal operation. Hence, to detect an attack using traditional anomaly detectors, the amplitude of the residual due to the attack is required to be sufficiently large such that $J(t) > J_{th}$ can be satisfied. Unfortunately, malicious adversaries may exploit this point to specially design attacks that generate residuals with sufficiently small amplitudes such that the attacks cannot trigger these anomaly detectors and remain undetected.

To distinguish the variables in the normal case (attack free), the superscript n is used. 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 denotes the change of x due to an attack, i.e., $x^a \triangleq x - x^n$. Now, we are ready to define the stealthy integrity attacks.

Definition 1. Consider an integrity attack $a(t)$ initiated at time T_0 and satisfying $a(t) \not\equiv 0$ for $t \geq T_0$. It is a stealthy integrity attack with respect to the anomaly detector $\mathcal{D}$ if

(a) $\|r^a(t)\| \to 0$ as $t \to +\infty$,

(b) $\|r(t)\|_{\mathrm{RMS}} \leq J_{th} - \delta$ for $t \geq T_0$, where $\delta > 0$ is a predefined scalar such that $J_{th} - \delta > 0$.

Remark 3. Comparing with the definition of perfect undetectable attacks in [21], Definition 1 is less restrictive but can capture the stealthiness characteristic. In particular, perfect undetectable attacks are not easily achieved by the attacker, sometimes impossible for general nonlinear CPS. Definition 1 alleviates this issue since conditions (a) and (b) are easier to achieve. It is worth pointing out that the attacks satisfying Definition 1 possess the same capabilities as those in [22] in terms of driving the system out of a safe region. ∇

III. Stealthy Integrity Attack Scenarios

An attack model to generate integrity attacks satisfying Definition 1 is proposed in this section. Two targets are achieved first: (i) proposing a more general form of attack models comparing to the ones in [3] and [11]; (ii) dealing with the issues arising due to the nonlinear function $g(t,x)$ and the nonlinear control law $u(t,\tilde{y},y_{\mathrm{ref}})$. Intuitively, regarding the more general attack model, an additionally additive input signal is introduced, providing a redundancy excitation. For dealing with the nonlinear g and u, we exploit several concepts such as incremental stability and controlled invariant subspace.

A. Stability of Incremental Systems

We start by deriving the incremental system of the detector $\mathcal{D}$ in the presence of an attack. To this end, $\tilde{u}^n$, $\tilde{u}$, $\tilde{y}^n$ and $\tilde{y}$ are explicitly given as follows:

$$\left\{\begin{array}{rcll} \tilde{u}^n(t) & = & u(t,\tilde{y}^n,y_{\mathrm{ref}}), & t < T_0, \\ \tilde{u}(t) & = & u(t,\tilde{y},y_{\mathrm{ref}}) + \Gamma_u a_u(t), & t \geq T_0, \end{array}\right. \tag{3}$$

$$\left\{\begin{array}{rcll} \tilde{y}^n(t) & = & Cx^n(t) + D_2 d(t), & t < T_0, \\ \tilde{y}(t) & = & Cx(t) + D_2 d(t) + \Gamma_y a_y(t), & t \geq T_0. \end{array}\right. \tag{4}$$

Let $u^a \triangleq u - u^n$. Then, from (1b), we have

$$u^a(t) = u(t,\tilde{y},y_{\mathrm{ref}}) - u(t,\tilde{y}^n,y_{\mathrm{ref}}). \tag{5}$$

It then follows from (3) that

$$\tilde{u}^a(t) \triangleq \tilde{u}(t) - \tilde{u}^n(t) = u^a(t) + \Gamma_u a_u(t). \tag{6}$$

From (4), $\tilde{y}^a \triangleq \tilde{y} - \tilde{y}^n$ due to an attack is obtained as

$$\tilde{y}^a(t) = Cx^a(t) + \Gamma_y a_y(t). \tag{7}$$

Thus, based on (2a), (2b), (5) and (7), the change of $\mathcal{D}$ due to an attack is described by

$$\mathcal{D}^a : \left\{\begin{array}{rcl} \dot{x}_r^a(t) & = & A_r x_r^a(t) + Bu^a(t) + K_r \tilde{y}^a(t) \\ & & +g(t,x_r) - g(t,x_r^n), \\ r^a(t) & = & Cx_r^a(t) - \tilde{y}^a(t), \end{array}\right. \tag{8}$$

where $x_r^a \triangleq x_r - x_r^n$ with $x_r^a(T_0) = 0$ and $r^a \triangleq r - r^n$. Then, the following lemma is given.

Lemma 1. *Under the i-ISS condition given in Assumption 1, $x_r^a(t)$ satisfies*

$$\|x_r^a(t)\| \leq \rho_1(\|u^a(t)\|) + \rho_2(\|\tilde{y}^a(t)\|),\ \forall\ t \geq T_0, \tag{9}$$

where ρ_1 and ρ_2 are specified in Assumption 1. ■

Proof. The system $\mathcal{D}^a$ in (8) is an incremental system between $\mathcal{D}$ and $\mathcal{D}^n$ regardless of x_r and x_r^n explicitly appearing in (8). Thus, from the i-ISS condition in Assumption 1, x_r^a satisfies

$$\begin{array}{rcl} \|x_r^a(t)\| & \leq & \beta_1(\|x_r^a(T_0)\|, t - T_0) \\ & & +\rho_1(\|u^a(t)\|) + \rho_2(\|\tilde{y}^a(t)\|),\ \forall\ t \geq T_0. \end{array}$$

Since $\|x_r^a(T_0)\| = 0$, then $\beta_1(\|x_r^a(T_0)\|, t - T_0) = 0$. Hence, the inequality (9) follows. □

Next, the incremental system of $\mathcal{W}$ in (1) is derived in the presence of an attack. From (1a), (1d), (5), (6) and (7), the changes of x and $\tilde{y}$ due to an attack can be written as

$$\mathcal{W}^a : \left\{\begin{array}{rcl} \dot{x}^a(t) & = & Ax^a(t) + g(t,x) - g(t,x^n) \\ & & +Bu^a(t) + B_a a(t), \\ \tilde{y}^a(t) & = & Cx^a(t) + D_a a(t), \end{array}\right. \tag{10}$$

where $x^a(T_0) = 0$, $B_a \triangleq [B\Gamma_u, 0_{n_x \times |K_y|}]$ and $D_a \triangleq [0_{n_y \times |K_u|}, \Gamma_y]$. In the sequel, a system realization of $\mathcal{W}^a$ is presented, which can split the nonlinear system $\mathcal{W}^a$ into two coupled systems: a linear system and a nonlinear system. To this end, the linear system $\mathcal{W}_1^a$ is given by

$$\mathcal{W}_1^a : \left\{\begin{array}{rcl} \dot{x}_1^a(t) & = & Ax_1^a(t) + B_a a(t), \\ \tilde{y}_1^a(t) & = & Cx_1^a(t) + D_a a(t), \end{array}\right. \tag{11}$$

and the nonlinear system $\mathcal{W}_2^a$ is given by

$$\mathcal{W}_2^a : \left\{\begin{array}{rcl} \dot{x}_2^a(t) & = & Ax_2^a(t) + g(t,x) - g(t,x^n) + Bu^a(t), \\ \tilde{y}_2^a(t) & = & Cx_2^a(t). \end{array}\right. \tag{12}$$

The initial condition of x_1^a and x_2^a are respectively chosen as

$$x_1^a(T_0) = z_0, \; x_2^a(T_0) = -z_0, \tag{13}$$

where $z_0 \in \mathbb{R}^{n_x}$ is any vector. The equality relationship between $\mathcal{W}^a$ and $\mathcal{W}_1^a \oplus \mathcal{W}_2^a$ (see the footnote[2]) are shown in the following lemma.

Lemma 2. *Consider the systems $\mathcal{W}^a$ in (10), $\mathcal{W}_1^a$ in (11) and $\mathcal{W}_2^a$ in (12). Then, $\mathcal{W}^a = \mathcal{W}_1^a \oplus \mathcal{W}_2^a$, i.e.,*

$$x^a(t) = x_1^a(t)+x_2^a(t), \; \tilde{y}^a(t) = \tilde{y}_1^a(t)+\tilde{y}_2^a(t), \; \forall\, t \geq T_0. \tag{14}$$

Moreover, in the context of Definition 1, the attack signal $a(t)$ in $\mathcal{W}^a$ is stealthy if $a(t)$ in $\mathcal{W}_1^a \oplus \mathcal{W}_2^a$ is stealthy. ■

Proof. From (10), (11) and (12), it follows that $\dot{x}_1^a + \dot{x}_2^a - \dot{x}^a = A(x_1^a + x_2^a - x^a)$ and $\tilde{y}_1^a + \tilde{y}_2^a - \tilde{y}^a = C(x_1^a + x_2^a - x^a)$, where, following from (13), the initial condition is $x_1^a(T_0) + x_2^a(T_0) - x^a(T_0) = 0$. Therefore, we obtain

$$x_1^a(t) + x_2^a(t) - x^a(t) = 0, \; \forall\; t \geq T_0,$$

which indicates that $\mathcal{W}^a = \mathcal{W}_1^a \oplus \mathcal{W}_2^a$, and also indicates that $\tilde{y}^a(t)$ has the same boundedness as $\tilde{y}_1^a(t) + \tilde{y}_2^a(t)$. Thus, in the context of Definition 1, if $a(t)$ in $\mathcal{W}_1^a \oplus \mathcal{W}_2^a$ is stealthy, then $a(t)$ in $\mathcal{W}^a$ is also stealthy. Hence, Lemma 2 is proved. □

The system $\mathcal{W}_2^a$ in (12) can have some convergence properties under the control u^a. An assumption on the convergence of x_2^a and on the boundedness of u^a is given as follows:

Assumption 2. (*Incremental Stability*) It is assumed that under the condition $\tilde{y}_1^a(t) \equiv 0$ for $t \geq T_0$ and $g(t,x) = g(t, x^n + x_2^a)$, there exists a class $\mathcal{KL}$ function β_2 such that the state x_2^a of $\mathcal{W}_2^a$ in (12) satisfies

$$\|x_2^a(t)\| \leq \beta_2 \left(\|z_0\|, t - T_0\right), \forall\; t \geq T_0. \tag{15}$$

Moreover, u^a in (5) satisfies $\|u^a(t)\| \leq \beta_u(\tilde{y}^a, t)$ where

$$\lim_{\tilde{y}^a \to 0, t \to \infty} \beta_u(\tilde{y}^a, t) = 0, \; \beta_u(0, t) = 0. \tag{16}$$

Remark 4. In the context of incremental systems, (15) in Assumption 2 is an incrementally asymptotic stability requirement (see the definition in [18]) for the system $\mathcal{W}$ in (1) in the nominal case. Specifically, in the case $\tilde{y}_1^a \equiv 0$ for $t \geq T_0$, it follows from (14) in Lemma 2 that $\tilde{y} = \tilde{y}^n + \tilde{y}_2^a$, and hence, it follows from (5) that $u^a(t) = u(t, \tilde{y}^n + \tilde{y}_2^a, y_{\rm ref}) - u(t, \tilde{y}^n, y_{\rm ref})$. Let $x_2 \triangleq x^n + x_2^a$ and $\tilde{y}_2 \triangleq \tilde{y}^n + \tilde{y}_2^a$. Then, since $g(t,x) = g(t, x^n + x_2^a) = g(t, x_2)$ and $u^a(t) = u(t, \tilde{y}_2, y_{\rm ref}) - u(t, \tilde{y}^n, y_{\rm ref})$, the differential equation in (12) is the incremental system between $\mathcal{W}_2 : \dot{x}_2(t) = Ax_2(t) + g(t, x_2) + Bu(t, \tilde{y}_2, y_{\rm ref})$ with $x_2(T_0) = x(T_0)$ and $\mathcal{W}^n : \dot{x}^n(t) = Ax^n(t) + g(t, x^n) + Bu(t, \tilde{y}^n, y_{\rm ref})$ with $x^n(T_0) = x(T_0) - z_0$, where $\mathcal{W}^n$ is the nominal system of $\mathcal{W}$ in (1). Based on the equivalence between asymptotic stability and incrementally asymptotic stability given in [18], a sufficient condition to guarantee (15) in Assumption 2 is that the system $\mathcal{W}^n$ is locally asymptotically stable. Therefore, we have that (15) in Assumption 2 is satisfied if in the nominal case and $y_{\rm ref}(t) \equiv 0$ for $t \geq 0$, the state x of $\mathcal{W}$ (or x^n of $\mathcal{W}^n$) can asymptotically converge to zero. Regarding (16) in Assumption 2, consider an example that u in (1) is a Lipschitz function of x. The existence of a β_u satisfying (16) follows straightforwardly. For more complex examples, the reader is referred to the **Case Study** section in this paper. ∇

Remark 5. The case that only the control inputs are attacked is a special case of the formulation in this paper, where $\Gamma_y = 0$ in (1). In this case, $\tilde{y}^a = y^a$ and $\tilde{y}^a$ is zero at T_0. Moreover, $\tilde{y}^a$ is required to remain low in amplitude for the stealthiness purpose. Under $\|u^a\| \leq \beta_u(\tilde{y}^a, t - T_0)$ and (16) in Assumption 2, u^a is also zero at T_0 and remains low in amplitude. However, a_u does not have to be identically zero or to remain low in amplitude if it is designed based on the methodology developed in this paper. The reason for this can be found from (6) and the explanation is that a zero u^a cannot lead to a zero a_u. Therefore, in the case that only the control inputs can be attacked, $a_y(t) \equiv 0$ for $t \geq T_0$, and the attack objective can be achieved by a_u alone. ∇

B. Largest Controlled Invariant Subspace

This section is to find the largest controlled invariant subspace such that in the presence of the input a, the nonlinear function $g(t,x)$ of $\mathcal{W}_2^a$ in (12) satisfies $g(t,x) = g(t, x^n + x_2^a)$, thereby being independent on x_1^a. Moreover, under the input a, the output $\tilde{y}_1^a$ of $\mathcal{W}_1^a$ in (11) is zero identically. We start by defining the largest controlled invariant subspace such that $g(t,x) = g(t, x^n + x_2^a)$. Inspired by the extended differential mean value theorem in [23], the deviation between $g(t,x)$ and $g(t, x^n + x_2^a)$ can be written as

$$g(t,x) - g(t, x^n + x_2^a) = G(t, \xi) x_1^a(t), \tag{17}$$

where $\xi \triangleq \Phi(x, x^n + x_2^a) = [\xi_1, \cdots, \xi_{n_x}] \in \mathbb{R}^{n_x \times n_x}$ with $\xi_i \in \mathrm{Co}(x, x^n + x_2^a)$ for $i = 1, \cdots, n_x$. Moreover,

$$G(t,\xi) \triangleq \begin{bmatrix} \frac{\partial g_1}{\partial x_1}(t,\xi_1) & \cdots & \frac{\partial g_1}{\partial x_{n_x}}(t,\xi_1) \\ \vdots & \ddots & \vdots \\ \frac{\partial g_{n_x}}{\partial x_1}(t,\xi_{n_x}) & \cdots & \frac{\partial g_{n_x}}{\partial x_{n_x}}(t,\xi_{n_x}) \end{bmatrix}, \tag{18}$$

where g_i is the i-th element of g, and x_i is the i-th element of x. It is worth pointing out that $\mathrm{Co}(x, x^n + x_2^a)$ denotes the convex hull of the set $\{x, x^n + x_2^a\}$, which indicates that $\xi = \Phi(x, x^n + x_2^a)$ takes an unknown specific value and hence is uncertain. A suitable way to guarantee $g(t,x) = g(t, x^n + x_2^a)$ is that in the presence of any uncertainty $\xi \in \mathbb{R}^{n_x \times n_x}$, x_1^a is decoupled with $G(t,\xi)$, i.e., $G(t,\xi) x_1^a(t) \equiv 0$ in (17) for $t \geq T_0$. In the sequel, the largest controlled invariant subspace decoupled with $G(t,\xi)$ is proposed. Let $\mathcal{H}$ be a constant linear subspace of $\ker(G(t,\xi))$. We then have the following lemma.

Lemma 3. *If there exist m indices $\{i_1, \cdots, i_m\} \subseteq \{1, \cdots, n_x\}$, $m \leq n_x$, and $m-1$ constant scalars $k_1, \cdots, k_{m-1}$ such that $G_{i_1} = k_1 G_{i_2} + \cdots + k_{m-1} G_{i_m}$ where G_j represents the j-th column of $G(t,\xi)$ in (18), then a nontrivial subspace[3] $\mathcal{H} \subset \mathbb{R}^{n_x}$ and $\mathcal{H} \neq \emptyset$ exists.* ■

[2] The operation $\mathcal{W}_1^a \oplus \mathcal{W}_2^a$ leads to a dynamical system with state $x_1^a + x_2^a$ and output $\tilde{y}_1^a + \tilde{y}_2^a$.

[3] In the vector space $\mathbb{R}^{n_x}$, $\{\mathbf{0}\}$ and $\mathbb{R}^{n_x}$ are the trivial subspace.

Proof. Based on linear algebra theory, there exists an elementary matrix determined by $i_1, \cdots, i_m$ and $k_1, \cdots, k_{m-1}$, denoted by $E(i_1, \cdots, i_m, k_1, \cdots, k_{m-1})$, such that $G(\cdot)E(\cdot) = [G_1, \cdots, G_{(i_1-1)}, 0, G_{(i_1+1)} \cdots, G_{n_x}]$. Then, $G(\cdot)E(\cdot)[0_1, \cdots, 0_{i-1}, 1, 0_{i+1}, \cdots, 0_{n_x}]^T = 0$. Thus, we can obtain $\mathcal{H} = \mathrm{Im}(E(\cdot)[0_1, \cdots, 0_{i-1}, 1, 0_{i+1}, \cdots, 0_{n_x}]^T)$. Hence, the result follows. □

Remark 6. Many classes of nonlinear functions $g(t,x)$ satisfy the condition in Lemma 3. Two intuitive classes are discussed: 1) $g(t,x)$ is independent on some components of x, for example, $g(t,x) = [x_2^2, x_2 \sin(x_2)]^T$ with $x = [x_1, x_2]^T$. In this example, $G_1 = 0$ and $\mathcal{H} = \mathrm{Im}[1,0]^T$; 2) there exist two elements x_i and x_j such that $\frac{\partial g(t,x)}{\partial x_i} = k \frac{\partial g(t,x)}{\partial x_j}$ with k being any scalar. For instance, $g(t,x) = [x_1x_3 + x_2x_3, x_1x_4 + x_2x_4, 0, 0]^T$ 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 hence $G_1 = G_2$. Thus, $\mathcal{H} = \mathrm{Im}[1,-1,0,0]^T$. ∇

Remark 7. From Lemma 3, any uncertainties in the columns $\{1, \cdots, n_x\}/\{i_1, \cdots, i_m\}$ of $G(t,\xi)$ do not affect the existence and the base vectors of $\mathcal{H}$. Therefore, some items of $g(t,x)$ relying on x_j with $j \in \{1, \cdots, n_x\}/\{i_1, \cdots, i_m\}$, possessing parameter uncertainties or being unknown by the attacker, may not affect the existence and the base vectors of $\mathcal{H}$. For example, consider $g'(t,x) = [x_1x_3 + x_2x_3, x_1x_4 + x_2x_4, 0, 0]^T$ with $x = [x_1, x_2, x_3, x_4]^T$ and an uncertain function $g(t,x) = [x_1x_3 + x_2x_3 + f(x_3,x_4), x_1x_4 + x_2x_4, 0, 0]^T$, where $f(x_3,x_4)$ is the uncertainty or an unknown function to the attacker. In this example, $n_x = 4$, $m = 2$, $i_1 = 1$ and $i_2 = 2$, and a common $\mathcal{H}$ of $g(t,x)$ and $g'(t,x)$ is $\mathcal{H} = \mathrm{Im}[1,-1,0,0]^T$. Therefore, the uncertainty $f(x_3,x_4)$ relying on x_3 and x_4 has no effects on the existence and the base vectors of $\mathcal{H}$, which proves the above analysis. ∇

For any subspace $\mathcal{H} \subset \mathbb{R}^{n_x}$, the largest controlled invariant subspace of $\mathcal{W}_1^a$ in (11) contained in $\mathcal{H}$ is defined as follows

$$\mathcal{V}(\mathcal{H}) \triangleq \{z_0 \in \mathbb{R}^{n_x} \mid \exists\ a(t), x_1^a(t; z_0, a) \in \mathcal{H},\ \forall\ t \geq T_0\},$$

where $x_1^a(t; z_0, a)$ is the solution to (11) with initial condition z_0 and input a. Based on Definition 2 in *Appendix* and [17], there exists a matrix F_a satisfying

$$(A + B_aF_a)\mathcal{V}(\mathcal{H}) \subset \mathcal{V}(\mathcal{H}). \tag{19}$$

On the other hand, according to [17], the weakly unobservable subspace $\mathcal{V}(\mathcal{W}_1^a)$ of $\mathcal{W}_1^a$ in (11) is the largest subspace such that there exists a matrix F_a satisfying

$$(A + B_aF_a)\mathcal{V}(\mathcal{W}_1^a) \subset \mathcal{V}(\mathcal{W}_1^a),\ (C + D_aF_a)\mathcal{V}(\mathcal{W}_1^a) = 0. \tag{20}$$

By combining (19) and (20), we have the following lemma.

Lemma 4. *(i) Let $\mathcal{V}_0 = \mathcal{V}(\mathcal{W}_1^a) \cap \mathcal{V}(\mathcal{H})$. Then, there exists a matrix F_a such that*

$$(A + B_aF_a)\mathcal{V}_0 \subset \mathcal{V}_0,\ (C + D_aF_a)\mathcal{V}_0 = 0. \tag{21}$$

(ii) Let $z_0 \in \mathcal{V}_0$, F_a satisfy (21) and L_a satisfy

$$\mathrm{Im}L_a = \ker D_a\ \cap\ B_a^{-1}\mathcal{V}_0. \tag{22}$$

Then, $x_1^a(t; z_0, a) \in \mathcal{V}_0$ if and only if $a(t)$ is designed as

$$a(t) = F_ax_1^a(t) + L_aw(t), \tag{23}$$

where $w(t)$ is any signal with proper dimensions. ■

Proof. *(i)* The result in (21) can be derived directly by the following equations: $(A+B_aF_a)\mathcal{V}_0 = (A+B_aF_a)\mathcal{V}(\mathcal{W}_1^a) \cap (A+B_aF_a)\mathcal{V}(\mathcal{H}) = \mathcal{V}(\mathcal{W}_1^a) \cap \mathcal{V}(\mathcal{H}) = \mathcal{V}_0$ and $(C + D_aF_a)\mathcal{V}_0 = (C + D_aF_a)\mathcal{V}(\mathcal{W}_1^a) \cap (C + D_aF_a)\mathcal{V}(\mathcal{H}) = 0$.
(ii) According to Theorem 2 given in the *Appendix* and given a $z_0 \in \mathcal{V}(\mathcal{H})$, $x_1^a(t; z_0, a) \in \mathcal{V}(\mathcal{H})$ for $t \geq T_0$ if and only if $a(t)$ is designed as the form (23) with F_a satisfying (19) and L_a satisfying $\mathrm{Im}L_a = B_a^{-1}\mathcal{V}(\mathcal{H})$. Also, Theorem 3 given in the *Appendix* proves that given a $z_0 \in \mathcal{V}(\mathcal{W}_1^a)$, $x_1^a(t; z_0, a) \in \mathcal{V}(\mathcal{W}_1^a)$ for $t \geq T_0$ if and only if $a(t)$ is designed as the form (23) with F_a satisfying (20) and L_a satisfying $\mathrm{Im}L_a = \ker D_a\ \cap\ B_a^{-1}\mathcal{V}(\mathcal{W}_1^a)$. Hence, the result (b) follows. □

C. Attack Model and Stealthiness Analysis

In this subsection, the attack model for generating attacks satisfy Definition 1 is designed. We start by showing the available resources to the attacker as follows:

1) *Model knowledge:* A, $B\Gamma_u$ (or B_a), C and a constant linear subspace $\mathcal{H}$ of $\ker(G)$ are known by the attacker;
2) *Disruption resource:* the sensor and actuator communication channels $K_y \subseteq \{1, \cdots, n_y\}$ and $K_u \subseteq \{1, \cdots, n_u\}$ can be compromised by the attacker.

Also, a geometric condition is needed and given as follows.

Assumption 3. For the nonlinear function $g(t,x)$ and the system $\mathcal{W}_1^a$ in (11), there exists a nontrivial $\mathcal{V}_0 = \mathcal{V}(\mathcal{W}_1^a) \cap \mathcal{V}(\mathcal{H})$ such that $\mathcal{V}_0 \neq \emptyset$.

Remark 8. Regarding the *model knowledge*, as analyzed in *Remark 7*, a nontrivial $\mathcal{H}$ can be obtained even through the attacker does not exactly know $g(t,x)$. The *disruption resource* indicates that the attacker does not to need to compromise all the sensor and actuator communication channels. The *model knowledge* in terms of $g(t,x)$ and *disruption resource* available to the attacker determine $\mathcal{V}(\mathcal{H})$ and $\mathcal{V}(\mathcal{W}_1^a)$ respectively, thereby affecting the existence of a nontrivial $\mathcal{V}_0$. ∇

Provided $\mathcal{W}_1^a$ in (11) and $\mathcal{W}_2^a$ in (12), the idea of designing the attack model is to exploit $\mathcal{V}_0$ such that $\tilde{y}_1^a(t) \equiv 0$ and $G(t,\xi)x_1^a(t) \equiv 0$ for $t \geq T_0$ and $\xi \in \mathbb{R}^{n_x \times n_x}$, and then to use u^a defined in (5) for stabilizing $\mathcal{W}_2^a$ based on the incremental stability. Following this idea, the attack model is proposed as

$$\mathcal{G}: \left\{ \begin{array}{rcl} \dot{z}(t) & = & (A + B_aF_a)z(t) + B_aL_aw(t), \\ a(t) & = & F_az(t) + L_aw(t), \end{array} \right. \tag{24}$$

where $z(T_0) = x_1^a(T_0) = z_0$ with $z_0 \in \mathcal{V}_0$, F_a is determined by (21), and L_a by (22). Moreover, $w(t)$ can be any signal vector with proper dimensions. The stealthiness of the generated attacks by (24) is analyzed in the following theorem.

Theorem 1. *Under Assumptions 1-3, there exists a compact region $\Omega_0(\delta) \subset \mathbb{R}^{n_x}$ containing the origin such that if*

$$z_0 \in \mathcal{V}_0 \cap \Omega_0(\delta), \tag{25}$$

where $\delta > 0$ specified in Definition 1, then the attack $a(t)$ generated by $\mathcal{G}$ is a stealthy attack satisfying Definition 1. ■

Proof. Following the sequence of the two conditions (a) and (b) shown in Definition 1, the proof is given as follows:

Step 1: Proving condition (a) in Definition 1. We start by writing $\mathcal{W}_1^a$ in (11) and $\mathcal{G}$ in (24) as an equivalent form in the coordinates $(\bar{x}_1^a, z)$, where $\bar{x}_1^a \triangleq x_1^a - z$, as follows:

$$\dot{\bar{x}}_1^a(t) = A\bar{x}_1^a(t),\ \dot{z}(t) = (A + B_aF_a)z(t) + B_aL_aw(t),$$
$$\tilde{y}_1^a(t) = C\bar{x}_1^a(t) + (C + D_aF_a)z(t) + D_aL_aw(t),$$

where $\bar{x}_1^a(T_0) = x_1^a(T_0) - z(T_0) = 0$. Since $\bar{x}_1^a(T_0) = 0$, $\bar{x}_1^a(t) \equiv 0$ for $t \geq T_0$, and hence, $x_1^a(t) = z(t)$ for $t \geq T_0$. Based on the result (ii) in Lemma 4, for $z_0 \in \mathcal{V}_0$, F_a satisfies (21), and L_a satisfies (22), we have $z(t) \in \mathcal{V}_0$, and hence, $G(t,\xi)z(t) \equiv 0$, $(C + D_aF_a)z(t) \equiv 0$ for all $t \geq T_0$. In addition, since L_a satisfies (22), $D_aL_aw(t) \equiv 0$ for all $t \geq T_0$. Thus, for all $t \geq T_0$ and all $\xi \in \mathbb{R}^{n_x \times n_x}$,

$$\tilde{y}_1^a(t) \equiv 0,\ G(t,\xi)x_1^a(t) = G(t,\xi)z(t) \equiv 0. \tag{26}$$

Now consider $\mathcal{W}_2^a$ in (12). From (14) and (26), we have $\tilde{y}^a(t) = \tilde{y}_2^a(t)$. Based on the incremental stability characterized by (15) in Assumption 2, we can induce that $x_2^a(t) \to 0$ as $t \to \infty$ and from $\tilde{y}_2^a = Cx_2^a$, we can obtain

$$\|\tilde{y}_2^a(t)\| \leq \|C\|\beta_2(\|z_0\|, t - T_0), \tag{27}$$

which indicates that $\tilde{y}_2^a(t) \to 0$, as $t \to \infty$. In addition, from $\|u^a(t)\| \leq \beta_u(\tilde{y}^a, t)$ and (16), we can obtain $u^a(t) \to 0$ as $t \to \infty$. We proceed by considering $\mathcal{D}^a$ in (8). Under Assumption 1, based on Lemma 1 and (27), by using the triangle inequality, $r^a(t)$ in (8) satisfies

$$\begin{aligned}\|r^a(t)\| &\leq \|C\| \cdot \|x_r^a(t)\| + \|\tilde{y}_2^a(t)\| \\ &= \|C\|\left(\rho_1(\|u^a\|) + \rho_2(\|\tilde{y}_2^a\|)\right) + \|\tilde{y}_2^a(t)\| \triangleq \bar{r}^a(t).\end{aligned} \tag{28}$$

Then, from $\tilde{y}_2^a(t) \to 0$ and $u^a(t) \to 0$, we can obtain that $\|\bar{r}^a(t)\| \to 0$ as $t \to \infty$. Thus, from $\|r^a(t)\| \leq \bar{r}^a(t)$, it follows that $\|r^a(t)\| \to 0$ as $t \to \infty$. Hence, the condition (a) of Definition 1 is proved.

Step 2: Proving condition (b) in Definition 1. Based on the triangle inequality, $\|r\|_{\mathrm{RMS}} = \|r^n + r^a\|_{\mathrm{RMS}} \leq \|r^n\|_{\mathrm{RMS}} + \|r^a\|_{\mathrm{RMS}}$. The compact region $\Omega_0(\delta)$ can be chosen as

$$\Omega_0(\delta) \triangleq \{z_0\,|\bar{r}^a(t) \leq J_{th} - \delta - \|r^n(t)\|_{\mathrm{RMS}},\ t \geq T_0\}, \tag{29}$$

where $\bar{r}^a(t)$ is given in (28) and δ is specified in Definition 1. Then, $\|r(t)\|_{\mathrm{RMS}} \leq J_{th} - \delta$ for $z_0 \in \Omega_0$ and $t \geq T_0$, and the condition (b) in Definition 1 is proved. Therefore, Definition 1 is satisfied, and the result follows. □

Remark 9. The scalar δ is selected by the attacker. A "smaller" Ω_0 is obtained if a "larger" δ is used in (29), but it may lead to a trivial Ω_0. However, the larger the used δ is, the "more stealthy" is the attack $a(t)$ generated by (24). This is because a "larger" δ leads to a "smaller" Ω_0 and z_0 belonging to Ω_0 is smaller in magnitude. Therefore, under the attack $a(t)$ based on such a smaller z_0, the residual r of the detector $\mathcal{D}$ is much smaller than the threshold J_{th} and hence, the attack is even more stealthy. ▽

Remark 10. In the attack generation scheme of [11], there is no input w. In such a generation scheme, in order to enlarge the destructiveness of an attack, z should be enlarged (this can be observed from $x_1^a(t) = z(t)$) by a "large" z_0. However, for the stealthiness purpose, z_0 should be sufficiently "small". Thus, a trade-off between the attack performance and stealthiness exists in the attack generation scheme in [3] and [11]. The additive input w proposed in the attack model $\mathcal{G}$ in (24) plays a key role since it provides an alternative excitation when z_0 is zero or close to the origin. By providing such an excitation source w, the attack model $\mathcal{G}$ is able to avoid this trade-off.▽

It should be pointed out that if $z_0 = 0$, then $L_aw(t)$ is the unique excitation for the attack model $\mathcal{G}$ in (24). The following corollary shows the perfect stealthiness properties of the generated attacks by $\mathcal{G}$ when $z_0 = 0$.

Corollary 1. *Under Assumptions 1-3, in the case $z_0 = 0$, the attack $a(t)$ generated by $\mathcal{G}$ in (24) is a stealthy attack satisfying Definition 1. Moreover, the residual $r(t)$ due to such an attack $a(t) \not\equiv 0$ does not change, i.e., $r^a(t) \equiv 0$ for $t \geq T_0$.* ■

Proof. This proof proceeds according to the proof process for Theorem 1. Based on Theorem 1, the attack $a(t)$ generated by $\mathcal{G}$ is a stealthy attack satisfying Definition 1. Moreover, note that when $z_0 = 0$, the class $\mathcal{KL}$ comparison function β_2 satisfies $\beta_2(\|z_0\|, t - T_0) = 0$. From (27), $\|u^a\| \leq \beta_u(\tilde{y}^a, t - T_0)$ and (16) in Assumption 2, we can derive that $\tilde{y}_2^a(t) \equiv 0$ and $u^a(t) \equiv 0$ for $t \geq T_0$. Thus, for the class $\mathcal{K}_\infty$ comparison functions ρ_1 and ρ_2 in (28), $\rho_1(\|u^a(t)\|) \equiv 0$ and $\rho_2(\|\tilde{y}_2^a(t)\|) \equiv 0$ for $t \geq T_0$, which indicates that $\bar{r}^a(t) \equiv 0$ for $t \geq T_0$. Therefore, it follows from (28) that $\|r^a(t)\| \leq \bar{r}^a(t) \equiv 0$ and hence, $\|r^a(t)\| \equiv 0$ for $t \geq T_0$. □

IV. Case Study

In this section, the longitudinal navigation model of an air breathing hypersonic vehicle is considered. Based on [24], the closed-loop longitudinal navigation dynamics can be written as $\mathcal{W}$ where $x = [x_1, x_2, x_3]^T$ with x_1, x_2 and x_3 are the altitude, attack angle and pitch rate, respectively. The system matrices and nonlinear function are given as follows:

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0.7586 & 1 \\ 0 & 2.6489 & -1.6197 \end{bmatrix} \times 10^{-6},\ B = \begin{bmatrix} 1.5 & 0 \\ 0 & 6.5333 \\ 0 & 0 \end{bmatrix} \times 10^{-4},$$
$$C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix},\ D_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 3.9937 & 0 & 0 \end{bmatrix} \times 10^{-5},\ D_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

In addition, $g(t,x) = [0, g_2(t,x), g_3(t,x)]^T$ where $g_2(t,x) = 3.6412 \times 10^{-9}\sin(x_2)$ and $g_3(t,x) = -3.0475 \times 10^{-4}x_2^2 - 4.8088 \times 10^{-5}x_2^2x_3 + 2.1334 \times 10^{-6}x_2x_3$. The reference altitude is $y_{\mathrm{ref}} = 1100000$ ft.

The output tracking control strategy u in (1) is constructed based on the integration control strategy in [25] for tracking a constant reference signal, and a standard static output feedback method. Specifically, an integration variable $\eta(t) = \int_0^t (S\tilde{y} - y_{\mathrm{ref}})dt$ with $S = [0, 1]$ is introduced. Then, η and the state x of $\mathcal{W}$ in (1) in the nominal case constitute an augmented system, i.e., $\dot{\eta} = SCx - y_{\mathrm{ref}}$ and $\dot{x} = Ax + g(t,x) + Bu$. The control law u is designed as $u = K_1\eta + K_2\tilde{y}$ where $\tilde{y} = Cx$, K_1 and K_2 are designed such that regardless of y_{ref}, the augmented system is asymptotically stable. By using the *LMI Toolbox*, K_1 and K_2 are determined and u is given by

$$u(t, \tilde{y}, y_{\mathrm{ref}}) = \begin{bmatrix} 0.67\times10^{-4} \\ 0 \\ 0 \end{bmatrix} \eta(t) + \begin{bmatrix} -0.12 & 0 \\ 0 & -2.69\times10^{6} \\ 0 & -517.29 \end{bmatrix} \tilde{y}(t).$$

Such a u can guarantee that $\mathcal{W}$ asymptotically tracks y_{ref}. Moreover, $u^a(t)$ defined in (5) can be obtained as $u^a(t) = K_1\eta^a(t) + K_2\tilde{y}^a(t)$ where $\eta^a(t)$ and $\tilde{y}^a(t)$ converge to zero as t goes to ∞. Therefore, β_u in Assumption 2 can be chosen as $\beta_u(\tilde{y}^a, t) = 0.67 \times 10^{-4}\|\eta^a(t)\| + 2.6955 \times 10^6\|\tilde{y}^a(t)\|$ so that (16) in 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 $\|g(t,x) - g(t,\hat{z})\| \leq 0.0183\|x - \hat{z}\|$. The gain K_r of the detector $\mathcal{D}$ is then calculated as

$$K_r = \begin{bmatrix} 5.7787 & 0 \\ 0 & 7.0202 \\ 0 & 5.9246 \end{bmatrix},$$

which can guarantee that the estimation error system is asymptotically stable in the absence of disturbances and hence Assumption 1 is satisfied. Furthermore, K_r can also guarantee that the $\mathcal{H}_\infty$ performance index from $d(t)$ to $r(t)$ is smaller than $\gamma = 2$. For the simulation purpose, the initial condition of x is given by $[100000 \text{ ft}, \pi/4, \pi/5]^T$, $d(t)$ is given by $[0.003\cos(3t+0.2), 180+20\sin(40t), 0.063\sin(10t+0.2)]^T$ and the time window T_w is set $T_w = 0.5$ s. Moreover, according to the robust threshold design approach based on $\mathcal{H}_\infty$ diagnosis observer in [19], the threshold is calculated as $J_{th} = 10\gamma \sup_{t\in\mathbb{R}_{\geq 0}}(d^T(t)d(t))^{\frac{1}{2}}/\sqrt{T_w} = 5.6569 \times 10^3$ (where 10 represents the time duration of the finite time-horizon detector $\mathcal{D}$). In addition, the scalar δ in Definition 1 is set as $\delta = 0.1 \times 10^3$.

Regarding the *model knowledge*, the attacker knows only the structure of $g(t,x)$, but do not know its specific parameters. Based on the structure of $g(t,x)$, the attacker can obtain

$$G(t,\xi) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & G_{22}(\xi_2) & 0 \\ 0 & G_{32}(\xi_3) & G_{33}(\xi_3) \end{bmatrix},$$

where $\xi = [\xi_1, \xi_2, \xi_3]$. However, all G_{22}, G_{32} and G_{33} are not exactly known by the attacker. Based on Lemma 3, the attacker can choose $\mathcal{H} = \text{Im}[1,0,0]^T$. In addition, regarding the *disruption resource*, all the sensor and actuator communication channels can be affected, i.e, $\Gamma_u = I_2$ and $\Gamma_y = I_2$. Using the *Geometric Approach Toolbox* provided in [26], it is calculated that $\mathcal{V}(\mathcal{H}) = \mathcal{V}(\mathcal{W}_1^a) = \text{Im}[1,0,0]^T$. Thus, $\mathcal{V}_0 = \mathcal{V}(\mathcal{H}) \cap \mathcal{V}(\mathcal{W}_1^a) \neq \emptyset$, and Assumption 3 is satisfied. Based on Theorem 1, the attacker choose $z_0 = [-2500, 0, 0]^T$,

$$F_a = \begin{bmatrix} -1.3333\times10^{-4} & 0 & 0 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \text{ and } L_a = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Therefore, the attack model $\mathcal{G}$ in (24) is constructed by using a band-limited white noise as the additive input $w(t)$.

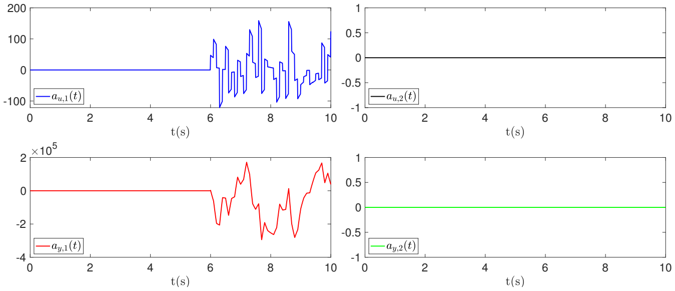

Fig. 2. The attack signals $a_u(t)$ and $a_y(t)$.

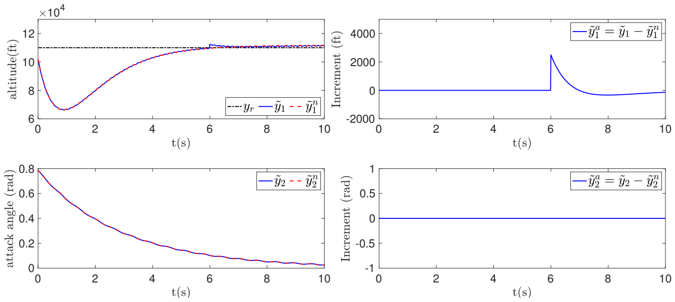

Fig. 3. The outputs $\tilde{y}$ and $\tilde{y}^n$ in the attack case (blue line) and non-attack case (red line), the reference signal y_{ref} (black line), and the change $\tilde{y}^a$.

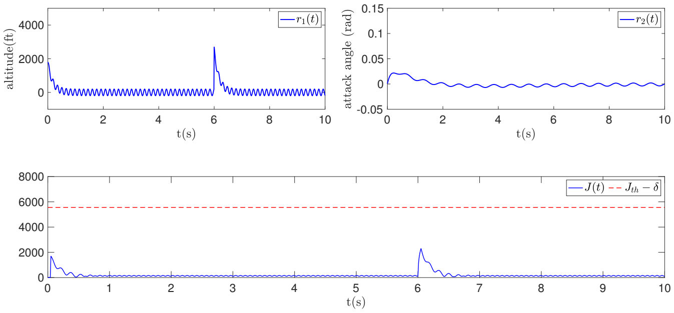

Fig. 4. The residual $r(t)$, the evaluation $J(t)$ and the threshold J_{th} of the anomaly detector $\mathcal{D}$.

The attack model $\mathcal{G}$ is activated at $T_0 = 6$s. The simulation results of the longitudinal navigation dynamics under the generated stealthy integrity attack are shown in Figs. 2-4. Figure 2 shows the time responses of the generated attack signals a_u and a_y respectively. As shown in Fig. 3, the altitude measurement $\tilde{y}_1$ under the attack a_u and a_y changes slightly, and Fig. 4 presents that the corresponding change $\tilde{y}_1^a$ cannot be detected by the anomaly detector $\mathcal{D}$ since $J(t) \leq J_{th} - \delta$ for $t > 6$s. Moreover, it can be observed from Fig. 3 that the change $\tilde{y}_1^a$ converges to zero asymptotically, and the change $\tilde{y}_2^a$ is identically zero. Consequently, this attack changes the altitude greatly but remains stealthy with respect to the anomaly detector $\mathcal{D}$.

We proceed to compare the result in this work with the ones in [3] and [11]. Since the attack generation schemes in [3] and [11] are for linear systems, linearization of the nonlinear longitudinal navigation model is needed. The linearization is done at the operation point $x_o = [100000, 0.01, 0]$, and the system matrix, denoted by A_o, is given as follows:

$$A_o = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 7.6225\times10^{-7} & 1 \\ 0 & 2.5880\times10^{-6} & -1.6195\times10^{-6} \end{bmatrix}.$$

Based on the design in [11], z_0 is chosen as $z_0 = [-2500, 3.5, 0]^T$. In the presence of the generated attack using attack model in [11], the detection result using $\mathcal{D}$ is shown in Fig. 5. It shows that $J(t) > J_{th}$ at about $t = 0.62$s and hence, the attack is detected by the anomaly detector $\mathcal{D}$, thereby being not stealthy. Therefore, the attacks generated based on [11] and according to the linearized system dynamics are not stealthy.

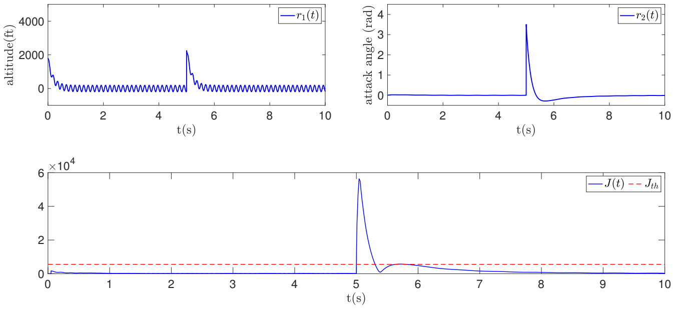

Fig. 5. The residual $r(t)$, the evaluation $J(t)$ and the threshold J_{th} of $\mathcal{D}$.

V. Conclusion

In this paper, a stealthy integrity attack generation methodology has been proposed for a class of nonlinear CPS. A type of stealthy integrity attacks has been defined for nonlinear CPS, and a more general attack model has been designed based on a controlled invariant subspace decoupled with the system outputs and the nonlinear function. The stealthiness has been rigorously investigated and a sufficient condition to guarantee the stealthiness was also derived. A simulation case study was presented to illustrate the effectiveness of the developed attack generation schemes. Future research efforts will be devoted to the detection of anomalies and stealthy integrity attacks and the distinction between them.

APPENDIX

Consider a linear time-invariant system Σ given as follows:

$$\Sigma : \begin{cases} \dot{x}(t) = & Ax(t) + Bu(t), \ x(0) = x_0, \\ y(t) = & Cx(t) + Du(t), \end{cases}$$

where x, y and u are the state, output and input, respectively.

Definition 2. [17] For the system Σ, a point x_0 is called weakly unobservable if there exists an input u such that the corresponding output satisfies $y(t; u, x_0) \equiv 0$ for $t \geq 0$. The set of all weakly unobservable points of Σ formulates the weakly unobservable subspace $\mathcal{V}(\Sigma)$ of Σ. ∎

Theorem 2. *(Theorem 4.3 in [17]) For the initial condition $x_0 \in \mathcal{V}(\Sigma)$, the state $x(t; u, x_0)$ resulting from u and x_0 remains belonging to $\mathcal{V}(\Sigma)$ if and only if $u(t) = Fx(t) + L\omega(t)$ where F and L satisfy $(A + BF)\mathcal{V}(\Sigma) \subset \mathcal{V}(\Sigma)$ and $\mathrm{Im}L = B^{-1}\mathcal{V}(\Sigma)$, respectively.* ∎

Theorem 3. *(Theorem 7.11 in [17]) For the initial condition $x_0 \in \mathcal{V}(\Sigma)$, the output $y(t; u, x_0) \equiv 0$ for $t \geq 0$ if and only if $u(t) = Fx(t) + L\omega(t)$ where F satisfies $(A + BF)\mathcal{V}(\Sigma) \subset \mathcal{V}(\Sigma)$, $(C + DF)\mathcal{V}(\Sigma) = 0$, L satisfies $\mathrm{Im}L = \ker D \cap B^{-1}\mathcal{V}(\Sigma)$, and $\omega(t)$ being any vector valued function with proper dimensions.* ∎

References

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

[2] F. Pasqualetti, F. Dorfler, and F. Bullo, "Control-theoretic methods for cyber physical security: Geometric principles for optimal cross-layer resilient control systems," *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 110–127, 2015.

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

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

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

[6] S. Dibaji, M. Pirani, D. Flamholz, A. Annaswamy, K. Johansson, and A. Chakrabortty, "A systems and control perspective of CPS security," *Annual Reviews in Control*, vol. 47, pp. 394–411, 2019.

[7] A. Gallo, M. Turan, F. Boem, T. Parisini, and G. Ferrari, "A distributed cyber-attack detection scheme with application to dc microgrids," *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3800–3815, 2020.

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

[9] R. Smith, "A decoupled feedback structure for covertly appropriating networked control systems," *IFAC Proceedings Volumes*, vol. 44, no. 1, pp. 90–95, 2011.

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

[11] A. Teixeira, I. Shames, H. Sandberg, and K. Johansson, "Revealing stealthy attacks in control systems," in *the 50th Annual Allerton Conference on Communication, Control, and Computing*. IEEE, 2012, pp. 1806–1813.

[12] S. Weerakkody, O. Ozel, P. Griffioen, and B. Sinopoli, "Active detection for exposing intelligent attacks in control systems," in *IEEE Conference on Control Technology and Applications*. IEEE, 2017, pp. 1306–1312.

[13] G. Park, C. Lee, H. Shim, Y. Eun, and K. Johansson, "Stealthy adversaries against uncertain cyber-physical systems: Threat of robust zero-dynamics attack," *IEEE Transactions on Automatic Control*, vol. 64, no. 12, pp. 4907–4919, 2019.

[14] Y. Mao, H. Jafarnejadsani, P. Zhao, E. Akyol, and N. Hovakimyan, "Novel stealthy attack and defense strategies for networked control systems," *IEEE Transactions on Automatic Control*, vol. 65, no. 9, 2020.

[15] G. Park, C. Lee, and H. Shim, "On stealthiness of zero-dynamics attacks against uncertain nonlinear systems: A case study with quadruple-tank process," in *International Symposium on Mathematical Theory of Networks and Systems (ISMTNS)*, 2018, pp. 10–17.

[16] 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. 13 682–13 687, 2020.

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

[18] D. Angeli, "A Lyapunov approach to incremental stability properties," *IEEE Transactions on Automatic Control*, vol. 47, no. 3, pp. 410–421, 2002.

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

[20] R. Rajamani, "Observers for Lipschitz nonlinear systems," *IEEE transactions on Automatic Control*, vol. 43, no. 3, pp. 397–401, 1998.

[21] J. Milošević, A. Teixeira, K. Johansson, and H. Sandberg, "Actuator security indices based on perfect undetectability: computation, robustness, and sensor placement," *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3816–3831, 2020.

[22] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 11, pp. 2715–2729, 2013.

[23] A. Zemouche, M. Boutayeb, and G. Bara, "Observer design for nonlinear systems: An approach based on the differential mean value theorem." in *the 44th IEEE Conference on Decision and Control*. IEEE, 2005, pp. 6353–6358.

[24] L. Fiorentini, A. Serrani, M. Bolender, and D. Doman, "Nonlinear robust adaptive control of flexible air-breathing hypersonic vehicles," *Journal of Guidance, Control, and Dynamics*, vol. 32, no. 2, pp. 402–417, 2009.

[25] D. Ye and G. Yang, "Adaptive fault-tolerant tracking control against actuator faults with application to flight control," *IEEE Transactions on control systems technology*, vol. 14, no. 6, pp. 1088–1096, 2006.

[26] G. Basile and G. Marro, *Controlled and conditioned invariants in linear system theory*. Prentice Hall Englewood Cliffs, NJ, 1992.