Resiliency of Nonlinear Control Systems to Stealthy Sensor Attacks

Amir Khazraei and Miroslav Pajic

***Abstract*— In this work, we focus on analyzing vulnerability of nonlinear dynamical control systems to stealthy sensor attacks. We define the notion of stealthy attacks in the most general form by leveraging Neyman-Pearson lemma. Specifically, an attack is considered to be stealthy if it is stealthy from (i.e., undetected by) any intrusion detector – i.e., the probability of the detection is not better than a random guess. We then provide a sufficient condition under which a nonlinear control system is vulnerable to stealthy attacks, in terms of moving the system to an unsafe region due to the attacks. In particular, we show that if the closed-loop system is incrementally exponentially stable while the open-loop plant is incrementally unstable, then the system is vulnerable to stealthy yet impactful attacks on sensors. Finally, we illustrate our results on a case study.**

I. Introduction

Cyber-physical systems (CPS) have been shown vulnerable to various types of cyber and physical attacks. Among adversarial attacks targeting CPSs, stealthy attacks designed by an intelligent attacker can have disastrous impact (e.g., [1]). Depending on the information available to the attacker, different types of stealthy attacks have been proposed. When only sensor measurements can be compromised, false data injection attacks are capable of significantly impacting the system while remaining undetected (i.e., stealthy) by a particular type of residual-based anomaly detectors (e.g., [2]–[8]). For example, for linear time invariant (LTI) systems, if measurements from all sensors can be compromised, the plant's (i.e., open-loop) instability is a necessary and sufficient condition for the existence of impactful stealthy attacks. Similarly, for LTI systems with strictly proper transfer functions, effective stealthy attacks on control input exist if the system has unstable zero invariant (e.g., [9], [10]). However, when the transfer function is not strictly proper, the attacker needs to compromise both plant's inputs and outputs. For such cases, e.g., [11] derives the conditions under which the system is vulnerable to stealthy attacks.

However, all these results have been shown only for LTI systems. Further, the notion of stealthiness is only characterized for a *specific type* of the employed intrusion detector (e.g., χ^2-based detectors). In [12], [13], the notion of attack stealthiness is generalized, defining an attack as stealthy if it is stealthy from the best existing intrusion detector; yet, the presented analysis only holds for LTI systems with LQG controllers. In addition, as we discuss in the paper, the notion of stealthiness defined in [12], [13] is time dependent and the probability of detection increases over time.

To the best of our knowledge, no existing work provides vulnerability analysis for systems with nonlinear dynamics, while considering general control and intrusion detector designs. In [14], covert attacks are introduced as stealthy attacks that can target a potentially nonlinear system. However, the attacker needs to have perfect knowledge of the system's dynamics and be able to compromise *both* the plant's input and outputs. More importantly, as the attack design is based on attacks on LTI systems, no guarantees are provided for effectiveness and stealthiness of attacks on nonlinear systems. Recently, [15] introduced stealthy attacks on a *specific class* of nonlinear systems with residual-based intrusion detector, but provided effective attacks only when *both* plant's inputs and outputs are compromised. On the other hand, in this work, we assume the attacker can only compromise the plant's sensing data and consider systems with *general* nonlinear dynamics. For systems with general nonlinear dynamics and residual-based intrusion detectors, machine learning-based attack design methods have been introduced (e.g., [16]), but without any theoretical analysis and guarantees regarding the impact of the stealthy attacks.

Consequently, in this work, we provide conditions for existence of effective yet stealthy attacks on nonlinear systems without limiting the analysis on a particular type of employed intrusion detectors. Our notion of attack stealthiness and system performance degradation is closely related to [17]. However, we extend these notions for systems with general nonlinear plants and controllers. Specifically, this is the first work considering the design of stealthy impactful sensor attacks for systems with general nonlinear dynamics that is independent of the deployed intrusion detector. Our main contributions are twofold. First, we introduce the notions of *strict* and ϵ-*stealthiness*. Second, using the well-known results for incremental stability from [18], we derive conditions for the existence of effective stealthy attacks that move the system into an unsafe operating region. We show that if the closed-loop system is incrementally stable while the open-loop plant is incrementally unstable, then the closed-loop system is strictly vulnerable to stealthy sensing attacks.

Notation: $\mathbb{R}, \mathbb{Z}, \mathbb{Z}_{\geq 0}$ denote the sets of reals, integers and non-negative integers, respectively, and $\mathbb{P}$ denotes the probability for a random variable. For a square matrix A, $\lambda_{max}(A)$ denotes the maximum eigenvalue. The p-norm of a vector x is $||x||_p$; when p is not specified, the 2-norm is implied. For a vector sequence, $x_0 : x_t$ denotes the set $\{x_0, x_1, ..., x_t\}$. A function $f : \mathbb{R}^n \to \mathbb{R}^p$ is Lipschitz with constant L if for any $x, y \in \mathbb{R}^n$ it holds that $||f(x) - f(y)|| \leq$

The authors are with the Department of Electrical & Computer Engineering, Duke University, Durham, NC 27708. Email: {amir.khazraei, miroslav.pajic}@duke.edu.

This work is sponsored by the ONR under agreement N00014-20-1-2745, AFOSR under the award number FA9550-19-1-0169, and by the NSF under CNS-1652544 award and the National AI Institute for Edge Computing Leveraging Next Generation Wireless Networks, Grant CNS-2112562.

$L||x-y||$. Finally, if $\mathbf{P}$ and $\mathbf{Q}$ are probability distributions relative to Lebesgue measure with densities $\mathbf{p}$ and $\mathbf{q}$, respectively, then the Kullback–Leibler (KL) divergence between $\mathbf{P}$ and $\mathbf{Q}$ is defined as $KL(\mathbf{P}||\mathbf{Q}) = \int \mathbf{p}(x) \log \frac{\mathbf{p}(x)}{\mathbf{q}(x)} dx$.

II. Preliminaries

Let $\mathbb{X} \subseteq \mathbb{R}^n$ and $\mathbb{D} \subseteq \mathbb{R}^m$, with $0 \in \mathbb{X}, \mathbb{D}$. Consider a discrete-time nonlinear system with an exogenous input, modeled in the state-space form as

$$x_{t+1} = f(x_t, d_t), \quad x_t \in \mathbb{X},\ t \in \mathbb{Z}_{\geq 0}, \tag{1}$$

where $f : \mathbb{X} \times \mathbb{D} \to \mathbb{X}$ is continuous and $f(0,0) = 0$. By $x(t, \xi, d)$, we denote the trajectory (i.e., the solution) of (1) at time t, when the system has the initial condition ξ and is subject to the input sequence $\{d_0 : d_{t-1}\}$; to simplify our notation, we denote the sequence $\{d_0 : d_{t-1}\}$ as d.

The following definitions are derived from [18]–[20].

Definition 1. *The system* (1) *is incrementally exponentially stable (IES) in the set $\mathbb{X} \subseteq \mathbb{R}^n$ if exist $\kappa > 1$ and $\lambda > 1$, that*

$$\|x(t, \xi_1, d) - x(t, \xi_2, d)\| \leq \kappa \|\xi_1 - \xi_2\| \lambda^{-t}, \tag{2}$$

holds for all $\xi_1, \xi_2 \in \mathbb{X}$, any $d_t \in \mathbb{D}$, and $t \in \mathbb{Z}_{\geq 0}$. When $\mathbb{X} = \mathbb{R}^n$, the system is referred to as globally incrementally exponentially stable (GIES).

Definition 2. *The system* (1) *is incrementally unstable (IU) in the set $\mathbb{X} \subseteq \mathbb{R}^n$ if for all $\xi_1 \in \mathbb{X}$ and any $d_t \in \mathbb{D}$, there exists a ξ_2 such that for any $M > 0$,*

$$\|x(t, \xi_1, d) - x(t, \xi_2, d)\| \geq M, \tag{3}$$

holds for all $t \geq t'$, for some $t' \in \mathbb{Z}_{\geq 0}$.

III. System Model

We now introduce the considered system and attack model, allowing us to formalize the problem addressed in this work.

A. System and Attack Model

We consider the setup from Figure 1 where each of the components is modeled as follows.

1) Plant: We assume the system evolves following a general nonlinear discrete-time dynamics in the state-space form

$$\begin{aligned} x_{t+1} &= f(x_t, u_t) + w_t, \\ y_t &= h(x_t) + v_t; \end{aligned} \tag{4}$$

here, $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $y \in \mathbb{R}^p$ are the state, input and output vectors of the plant, respectively. In addition, f is a nonlinear mapping from previous time state and control input to the current state, and h is the mapping from the states to the sensor measurements; we assume here that h is Lipschitz with a constant L_h. The plant output vector captures measurements from the set of plant sensors $\mathcal{S}$. Further, $w \in \mathbb{R}^n$ and $v \in \mathbb{R}^p$ are the process and measurement noises that are assumed to be Gaussian with zero mean, and Σ_w and Σ_v covariance matrices, respectively.

As we show later, it will be useful to consider the input to state relation of the dynamics (4); if we define $U = \begin{bmatrix} u^T & w^T \end{bmatrix}^T$, the first equation in (4) becomes

$$x_{t+1} = f_u(x_t, U_t). \tag{5}$$

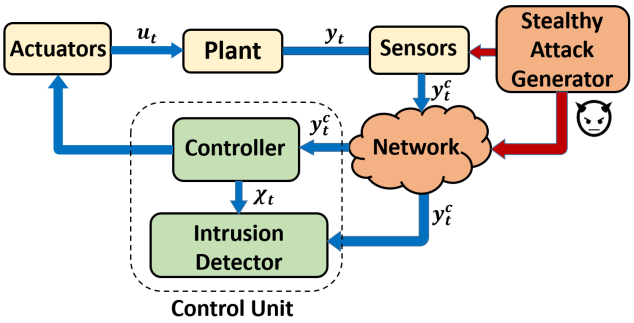

Fig. 1: Control system architecture considered in this work, in the presence of network-based attacks.

2) Control Unit: The controller, illustrated in Figure 1, is equipped with a feedback controller in the most general form, as well as an intrusion detector (ID). In what follows, we provide more details on the controller design. Intrusion detector will be discussed after introducing the attack model.

Controller: A large number of dynamical systems are intrinsically unstable or are designed to be unstable (e.g., if an aircraft is unstable, it is easier to change its altitude), and must be stabilized using a proper controller. Due to their robustness to uncertainties, closed-loop controllers are mainly used. In the most general form, a feedback controller design can be captured in the state-space form as

$$\begin{aligned} \chi_t &= f_c(\chi_{t-1}, y_t^c), \\ u_t &= h_c(\chi_t, y_t^c), \end{aligned} \tag{6}$$

where χ is the internal state of the controller, and y^c captures the sensor measurements received by the controller. Thus, without malicious activity, it holds that $y^c = y$; we assume that the communication network is reliable (e.g., wired). Note that the control model (6) is quite general, capturing for instance nonlinear filtering followed by a classic nonlinear controller (e.g., f_c can model an extended Kalman filter and h_c any full-state feedback controller).

We define the full state of the closed-loop system as $\mathbf{X} \triangleq \begin{bmatrix} x^T & \chi^T \end{bmatrix}^T$, and exogenous disturbances as $\mathbf{W} \triangleq \begin{bmatrix} w^T & v^T \end{bmatrix}$; then, the dynamics of the closed-loop system can be captured as

$$\mathbf{X}_{t+1} = F(\mathbf{X}_t, \mathbf{W}_t). \tag{7}$$

We assume that $\mathbf{X} = 0$ is the operating point of the noiseless system (i.e., when $w = v = 0$). Moreover, we assume f_c and h_c are designed to keep the system within a safe region around the equilibrium point. Here, without loss of generality, we define the safe region as $\mathbf{S} = \{x \in \mathbb{R}^n \mid \|x\|_2 \leq R_{\mathbf{S}}\}$, for some $R_{\mathbf{S}} > 0$.

3) Attack Model: We consider a sensor attack model where, for sensors from the set $\mathcal{K} \subseteq \mathcal{S}$, the information delivered to the controller differs from the non-compromised sensor measurements. The attacker can achieve this via e.g., noninvasive attacks such sensor spoofing (e.g., [21]) or by compromising information-flow from the sensors in $\mathcal{K}$ to the controller (e.g., as in network-based attacks [22]). In either

case, the attacker can launch false-date injection attacks, inserting a desired value instead of the current measurement of a compromised sensor.[1]

Thus, assuming that the attack starts at time $t = 0$, the sensor measurements delivered to the controller for $t \in \mathbb{Z}_{\geq 0}$ can be modeled as [23]

$$y_t^{c,a} = y_t^a + a_t; \tag{8}$$

here, $a_t \in \mathbb{R}^p$ denotes the attack signal injected by the attacker at time t via the compromised sensors from $\mathcal{K}$, y_t^a is the true sensing information (i.e., before the attack is injected at time t). In the rest of the paper, we assume $\mathcal{K} = \mathcal{S}$.

Since the controller uses the received sensing data to compute the input u_t, the compromised sensor values affect the evolution of the system and controller states. Hence, we add the superscript a to denote any signal obtained from a compromised system – e.g., thus, y_t^a is used to denote before-attack sensor measurements when the system is under attack in (8), and we denote the closed-loop plant and controller state when the system is compromised as $\mathbf{X}^a \triangleq \begin{bmatrix} x^a \\ \mathcal{X}^a \end{bmatrix}$.

In this work, we consider the commonly adopted threat model as in majority of existing stealthy attack designs, e.g., [2], [3], [5], [14], [24], where the attacker has full knowledge of the system, its dynamics and employed architecture. In addition, the attacker has the required computational power to calculate suitable attack signals to be injected, while planning ahead as needed. Finally, the attacker's goal is to design an attack signal a_t, $t \in \mathbb{Z}_{\geq 0}$, such that it always remains *stealthy* – i.e., undetected by the intrusion detection system – while *maximizing control performance degradation*. The notions of *stealthiness* and *control performance degradation* depend on the employed control architecture, and thus will be formally defined after the controller and intrusion detection have been introduced.

4) Intrusion Detector: To detect attacks (and anomalies), we assume that an ID is employed, analyzing the received sensor measurements and the internal controller state. Specifically, by defining $Y \triangleq \begin{bmatrix} y^c \\ \mathcal{X} \end{bmatrix}$, as well as $Y^a \triangleq \begin{bmatrix} y^{c,a} \\ \mathcal{X}^a \end{bmatrix}$ when the system is under attack, we assume that the ID has access to a sequence of values $Y_{-\infty} : Y_t$ until time t and solves the binary hypothesis checking problem

H_0: normal condition (the ID receives $Y_{-\infty} : Y_t$);
H_1: abnormal behaviour (receives $Y_{-\infty} : Y_{-1}, Y_0^a : Y_t^a$).[2]

Given a sequence of received data denoted by $\bar{Y}^t = \bar{Y}_{-\infty} : \bar{Y}_t$, it is either extracted from the distribution of the null hypothesis H_0, which we refer to as $\mathbf{P}$, or from an ***unknown*** distribution of the alternative hypothesis H_1, which we denote as $\mathbf{Q}$. Note that the *unknown* distribution $\mathbf{Q}$ is controlled by the attacker – i.e., the injected false data.

[1]We refer to sensors from $\mathcal{K}$ as compromised, even if a sensor itself is not directly compromised but its measurements may be altered due to e.g., network-based attacks.

[2]Since the attack starts at $t = 0$, we do not use superscript a for the system evolution for $t < 0$, as the trajectories of the non-compromised and compromised systems do not differ before the attack starts.

For a given ID mapping $D : \bar{Y}^t \to \{0, 1\}$, let us define $p_t^{TD}(D) = \mathbb{P}(D(\bar{Y}^t) = 1 | \bar{Y}^t \sim \mathbf{Q})$ as the probability of true detection, and $p_t^{FA}(D) = \mathbb{P}(D(\bar{Y}^t) = 1 | \bar{Y}^t \sim \mathbf{P})$ as the probability of false alarm for the detector D. We say that an ID (defined by D) to be better than a random guess-based ID (defined by D_{RG}) if $p^{FA}(D) < p^{TD}(D)$; the reason is that with the random guess ID it holds that $p^{FA}(D_{RG}) = \mathbb{P}(D_{RG}(\bar{Y}^t) = 1 | \bar{Y}^t \sim \mathbf{P}) = \mathbb{P}(D_{RG}(\bar{Y}^t) = 1) = \mathbb{P}(D_{RG}(\bar{Y}^t) = 1 | \bar{Y}^t \sim \mathbf{Q}) = p^{TD}(D_{RG})$.

IV. Formalizing Stealthy Attacks Requirements

In this section, we capture the conditions for which an attack sequence is stealthy even from an optimal ID. We define an attack to be stealthy if the best strategy for the ID is to ignore the measurements and make a random guess between the hypotheses; i.e., that there is ***no*** ID D that satisfies $p^{TD}(D) > p^{FA}(D)$. However, reaching such stealthiness guarantees may not be possible in general. Therefore, in addition to the notion of *strict stealthiness*, we define the notion of ϵ-*stealthiness*, which as we will show later, is attainable for a large class of nonlinear systems.

Definition 3. *Consider the system* (4)*. An attack sequence is* **strictly stealthy** *if there exists no detector for which $p_t^{FA} < p_t^{TD}$ holds, for any $t \geq 0$. An attack is* **ϵ-stealthy** *if for a given $\epsilon > 0$, there exists no detector such that $p_t^{FA} < p_t^{TD} - \epsilon$ holds, for any $t \geq 0$.*

Now, we can capture stealthiness conditions in terms of KL divergence of the corresponding distributions [25].

Theorem 1 ([25])**.** *An attack sequence is*

- *strictly stealthy if and only if $KL\big(\mathbf{Q}(Y_0^a : Y_t^a) || \mathbf{P}(Y_0 : Y_t)\big) = 0$ for all $t \in \mathbb{Z}_{\geq 0}$, where KL represents the Kullback–Leibler divergence operator.*
- *is ϵ-stealthy if the corresponding observation sequence $Y_0^a : Y_t^a$ satisfies*

$$KL\big(\mathbf{Q}(Y_0^a : Y_t^a) || \mathbf{P}(Y_0 : Y_t)\big) \leq \log(\frac{1}{1 - \epsilon^2}). \tag{9}$$

Remark 1. *The ϵ-stealthiness from [12], [13] requires*

$$\lim_{t \to \infty} \frac{KL\big(\mathbf{Q}(Y_0^a : Y_t^a) || \mathbf{P}(Y_0 : Y_t)\big)}{t} \leq \epsilon.$$

This allows the KL divergence to linearly increase over time for any $\epsilon > 0$; thus, after large-enough time period the attack may be detected. On the other hand, our definition of ϵ-stealthy only depends on ϵ and is fixed for any time t. Hence, it introduces a stronger notion of stealthiness for the attack.

A. Formalizing Attack Goal

As discussed, the attacker intends to *maximize* control performance degradation. As we consider the origin as the operating point, we formalize the attack objective as *maximizing (the norm of) the states x_t*; i.e., moving the system's states into an unsafe region. Since there might be a zone between the safe and unsafe region, we define the unsafe region as $\mathbf{U} = \{x \in \mathbb{R}^n \mid \|x\|_2 \geq \alpha\}$ for some $\alpha > R_{\mathbf{S}}$, where $R_{\mathbf{S}}$ is the radius of the safe region $\mathbf{S}$. Moreover, the

attacker wants *to remain stealthy (i.e., undetected by the intrusion detector)*, as formalized below.

Definition 4. *The attack sequence $\{a_0, a_1, ...\}$ is referred to as (ϵ, α)-successful attack if there exists $t' \in \mathbb{Z}_{\geq 0}$ such that $\|x^a_{t'}\| \geq \alpha$ and the attack is ϵ-stealthy for all $t \in \mathbb{Z}_{\geq 0}$. When such a sequence exists for a system, the system is called (ϵ, α)-attackable. When the system is (ϵ, α)-attackable for arbitrarily large α, it is referred to as perfectly attackable.*

Now, the problem considered in this work can be formalized as capturing the potential impact of stealthy attacks on a considered system; specifically, in the next section, we derive conditions for existence of a *stealthy* yet *effective* attack sequence $a_0, a_1, ...$ resulting in $\|x^a_t\| \geq \alpha$ for some $t \in \mathbb{Z}_{\geq 0}$ – i.e., we find conditions for the system to be (ϵ, α)-attackable. Here, for an attack to be stealthy, we focus on the $\epsilon-$stealthy notion; i.e., that even the best ID could only improve the detection probability by ϵ compared to the random-guess baseline detector.

V. Vulnerability Analysis of Nonlinear Systems to Stealthy Attacks

In this section, we derive the conditions such that the nonlinear system (4) with closed-loop dynamics (7) is vulnerable to effective stealthy attacks formally defined in Section IV. The following theorem captures such condition.

Theorem 2. *The system* (4) *is (ϵ, α)-attackable for arbitrarily large α and arbitrarily small ϵ, if the closed-loop dynamics* (7) *is incrementally exponentially stable (IES) in the set* $\mathbf{S}$ *and the system* (5) *is incrementally unstable (IU) in the set* $\mathbf{S}$.

Proof. Assume that the trajectory of the system and controller states for $t \in \mathbb{Z}_{<0}$ is denoted by $\mathbf{X}_{-\infty} : \mathbf{X}_{-1}$. Following the attack start at $t = 0$, let us consider the evolutions of the system with and without attacks for $t \in \mathbb{Z}_{\geq 0}$. For the system under attack, starting at time zero, the trajectory $\mathbf{X}^a_0 : \mathbf{X}^a_t$ of the system and controller states is governed by

$$\begin{aligned} x^a_{t+1} =& f(x^a_t, u^a_t) + w_t, \quad y^{c,a}_t = h(x^a_t) + v_t + a_t, \\ \mathcal{X}^a_t =& f_c(\mathcal{X}^a_{t-1}, y^{c,a}_t), \quad u^a_t = h_c(\mathcal{X}^a_t, y^{c,a}_t). \end{aligned} \tag{10}$$

On the other hand, if the system were not under attack during $t \in \mathbb{Z}_{\geq 0}$, we denote the plant and controller state evolution by $\mathbf{X}_0 : \mathbf{X}_t$. Hence, it is a continuation of the system trajectories $\mathbf{X}_{-\infty} : \mathbf{X}_{-1}$ if hypothetically no data-injection attack occurs during $t \in \mathbb{Z}_{\geq 0}$. Since the system and measurement noises are independent of the state, we can assume that $w^a_t = w_t$ and $v^a_t = v_t$. In this case, the dynamics of the plant and controller state evolution satisfies

$$\begin{aligned} x_{t+1} =& f(x_t, u_t) + w_t, \quad y^c_t = h(x_t) + v_t, \\ \mathcal{X}_t =& f_c(\mathcal{X}_{t-1}, y^c_t), \quad u_t = h_c(\mathcal{X}_t, y^c_t), \end{aligned} \tag{11}$$

which can be captured in the compact form (7), with $\mathbf{X}_0 = \begin{bmatrix} x_0^T & \mathcal{X}_0^T \end{bmatrix}^T$. Now, consider the sequence of attack vectors injected in the system from (10), which are constructed using the following dynamical model

$$\begin{aligned} s_{t+1} &= f(x^a_t, u^a_t) - f(x^a_t - s_t, u^a_t) \\ a_t &= h(x^a_t - s_t) - h(x^a_t), \end{aligned} \tag{12}$$

for $t \in \mathbb{Z}_{\geq 0}$, and with some arbitrarily chosen nonzero initial value of s_0. By injecting the above attack sequence into the sensor measurements, we can verify that $y^{c,a}_t = h(x^a_t) + v_t + a_t = h(x^a_t - s_t) + v_t$. After defining $e_t \triangleq x^a_t - s_t$ and combining (12) with (10), the dynamics of e_t and the controller, and the corresponding input and output satisfy

$$\begin{aligned} e_{t+1} =& f(e_t, u^a_t) + w_t, \quad y^{c,a}_t = h(e_t) + v_t, \\ \mathcal{X}^a_t =& f_c(\mathcal{X}^a_{t-1}, y^{c,a}_t), \quad u^a_t = h_c(\mathcal{X}^a_t, y^{c,a}_t), \end{aligned} \tag{13}$$

with the initial condition $e_0 = x^a_0 - s_0$.

Now, if we define $\mathbf{X}^e_t = \begin{bmatrix} e_t \\ \mathcal{X}^a_t \end{bmatrix}$, it holds that

$$\mathbf{X}^e_{t+1} = F(\mathbf{X}^e_t, \mathbf{W}_t), \tag{14}$$

with $\mathbf{X}^e_0 = \begin{bmatrix} e_0 \\ \mathcal{X}^a_0 \end{bmatrix}$. Since we have that $x^a_0 = x_0$ and $\mathcal{X}^a_0 = \mathcal{X}_0$, it holds that $\mathbf{X}_0 - \mathbf{X}^e_0 = \begin{bmatrix} s_0 \\ 0 \end{bmatrix}$. On the other hand, since both (14) and (7) share the same function and argument $\mathbf{W}_t$, the closed-loop system (14) is IES, and it also follows that

$$\begin{aligned} \|\mathbf{X}(t, \mathbf{X}_0, \mathbf{W}) - \mathbf{X}^e(t, \mathbf{X}^e_0, \mathbf{W})\| &\leq \kappa \|\mathbf{X}_0 - \mathbf{X}^e_0\| \lambda^{-t} \\ &\leq \kappa \|s_0\| \lambda^{-t}; \end{aligned} \tag{15}$$

therefore, the trajectories of $\mathbf{X}$ (i.e., the system without attack) and $\mathbf{X}^e$ converge to each other exponentially fast.

We now use these results to show that the generated attack sequence satisfies the ϵ-stealthiness condition. By defining $\mathbf{Z}_t = \begin{bmatrix} x_t \\ y^c_t \end{bmatrix}$ and $\mathbf{Z}^e_t = \begin{bmatrix} e_t \\ y^{c,a}_t \end{bmatrix}$, it holds that

$$\begin{aligned} & KL\big(\mathbf{Q}(Y^a_0 : Y^a_t) || \mathbf{P}(Y_0 : Y_t)\big) \\ & \overset{(i)}{\leq} KL\big(\mathbf{Q}(\mathbf{X}^e_0 : \mathbf{X}^e_t) || \mathbf{P}(\mathbf{X}_0 : \mathbf{X}_t)\big) \\ & \overset{(ii)}{\leq} KL\big(\mathbf{Q}(\mathbf{Z}_{-\infty} : \mathbf{Z}_{-1}, \mathbf{Z}^e_0 : \mathbf{Z}^e_t) || \mathbf{P}(\mathbf{Z}_{-\infty} : \mathbf{Z}_{-1}, \mathbf{Z}_0 : \mathbf{Z}_t)\big), \end{aligned} \tag{16}$$

where we applied the data-processing inequality property of KL-divergence for $t \in \mathbb{Z}_{\geq 0}$ to obtain (i), and the monotonicity property of KL-divergence to obtain the inequality (ii).[3] Then, we apply the chain-rule property of KL-divergence on the right-hand side of (16) to obtain the following

$$\begin{aligned} & KL\big(\mathbf{Q}(\mathbf{Z}_{-\infty} : \mathbf{Z}_{-1}, \mathbf{Z}^e_0 : \mathbf{Z}^e_t) || \mathbf{P}(\mathbf{Z}_{-\infty} : \mathbf{Z}_{-1}, \mathbf{Z}_0 : \mathbf{Z}_t)\big) \\ & = KL\big(\mathbf{Q}(\mathbf{Z}_{-\infty} : \mathbf{Z}_{-1}) || \mathbf{P}(\mathbf{Z}_{-\infty} : \mathbf{Z}_{-1})\big) + \\ & \quad KL\big(\mathbf{Q}(\mathbf{Z}^e_0 : \mathbf{Z}^e_t | \mathbf{Z}_{-\infty} : \mathbf{Z}_{-1}) || \mathbf{P}(\mathbf{Z}_0 : \mathbf{Z}_t | \mathbf{Z}_{-\infty} : \mathbf{Z}_{-1})\big) \\ & = KL\big(\mathbf{Q}(\mathbf{Z}^e_0 : \mathbf{Z}^e_t | \mathbf{Z}_{-\infty} : \mathbf{Z}_{-1}) || \mathbf{P}(\mathbf{Z}_0 : \mathbf{Z}_t | \mathbf{Z}_{-\infty} : \mathbf{Z}_{-1})\big); \end{aligned} \tag{17}$$

here, we used the fact that the KL-divergence of two identical distributions (i.e., $\mathbf{Q}(\mathbf{Z}_{-\infty} : \mathbf{Z}_{-1})$ and $\mathbf{P}(\mathbf{Z}_{-\infty} : \mathbf{Z}_{-1})$ since the system is not under attack for $t < 0$) is zero.

[3] Due to the space limitation, we do not introduce data-processing, chain-rule, and monotonicity properties of KL-divergence. More information about these terms can be found in [26].

Applying the chain-rule property of KL-divergence to (17) results in

$$\begin{aligned}
&KL\big(\mathbf{Q}(\mathbf{Z}_0^e:\mathbf{Z}_t^e|\mathbf{Z}_{-\infty}:\mathbf{Z}_{-1})||\mathbf{P}(\mathbf{Z}_0:\mathbf{Z}_t|\mathbf{Z}_{-\infty}:\mathbf{Z}_{-1})\big)\\
&\quad\le KL\big(\mathbf{Q}(e_0|\mathbf{Z}_{-\infty}:\mathbf{Z}_{-1})||\mathbf{P}(x_0|\mathbf{Z}_{-\infty}:\mathbf{Z}_{-1})\big)\\
&\quad+KL\big(\mathbf{Q}(y_0^{c,a}|e_0,\mathbf{Z}_{-\infty}:\mathbf{Z}_{-1})||\mathbf{P}(y_0|x_0,\mathbf{Z}_{-\infty}:\mathbf{Z}_{-1})\big)\\
&\quad+...+KL\big(\mathbf{Q}(e_t|\mathbf{Z}_{-\infty}:\mathbf{Z}_{t-1}^e)||\mathbf{P}(x_t|\mathbf{Z}_{-\infty}:\mathbf{Z}_{t-1})\big)\\
&\quad+KL\big(\mathbf{Q}(y_t^{c,a}|e_t,\mathbf{Z}_{-\infty}:\mathbf{Z}_{t-1}^e)||\mathbf{P}(y_t|x_t,\mathbf{Z}_{-\infty}:\mathbf{Z}_{t-1})\big).
\end{aligned}\tag{18}$$

Given $\mathbf{Z}_{-\infty}:\mathbf{Z}_{t-1}$, the distribution of x_t is a Gaussian with mean $f(x_{t-1},u_{t-1})$ and covariance Σ_w. Similarly given $\mathbf{Z}_{-\infty}:\mathbf{Z}_{-1},\mathbf{Z}_0^e:\mathbf{Z}_{t-1}^e$, the distribution of e_t is a Gaussian with mean $f(e_{t-1},u_{t-1}^a)$ and covariance Σ_w. Since we have that $x_t=f(x_{t-1},u_{t-1})+w_t$ and $e_t=f(e_{t-1},u_{t-1}^a)+w_t$ according to (11) and (13), it holds that $f(x_{t-1},u_{t-1})-f(e_{t-1},u_{t-1}^a)=x_t-e_t$. On the other hand, in (15) we showed that $\|x_t-e_t\|\le\kappa\|s_0\|\lambda^{-t}$ holds for $t\in\mathbb{Z}_{\ge0}$. Therefore, for all $t\in\mathbb{Z}_{\ge0}$, it holds that

$$\begin{aligned}
&KL\big(\mathbf{Q}(e_t|\mathbf{Z}_{-\infty}:\mathbf{Z}_{t-1}^e)||\mathbf{P}(x_t|\mathbf{Z}_{-\infty}:\mathbf{Z}_{t-1})\big)=\\
&\quad=(x_t-e_t)^T\Sigma_w^{-1}(x_t-e_t)\le\kappa^2\|s_0\|^2\lambda^{-2t}\lambda_{max}(\Sigma_w^{-1}),
\end{aligned}\tag{19}$$

where $\lambda_{max}(\Sigma_w^{-1})$ is the maximum eigenvalue of Σ_w^{-1}.

Now, using the Markov property it holds that $\mathbf{Q}(y_t^{c,a}|e_t,\mathbf{Z}_{-\infty}:\mathbf{Z}_{t-1}^e)=\mathbf{Q}(y_t^{c,a}|e_t)$ and $\mathbf{P}(y_t|x_t,\mathbf{Z}_{-\infty}:\mathbf{Z}_{t-1})=\mathbf{P}(y_t|x_t)$; also, from (11) and (13) it holds that given x_t and e_t, $\mathbf{P}(y_t|x_t)$ and $\mathbf{Q}(y_t^{c,a}|e_t)$ are both Gaussian with mean $h(x_t)$ and $h(e_t)$, respectively, and covariance Σ_v. Thus, it follows that

$$\begin{aligned}
&KL\big(\mathbf{Q}(y_t^{c,a}|e_t)||\mathbf{P}(y_t|x_t)\big)\\
&=\big(h(x_t)-h(e_t)\big)^T\Sigma_v^{-1}\big(h(x_t)-h(e_t)\big)\\
&\le L_h^2(x_t-e_t)^T\Sigma_v^{-1}(x_t-e_t)\le L_h^2\kappa^2\|s_0\|^2\lambda^{-2t}\lambda_{max}(\Sigma_v^{-1}).
\end{aligned}\tag{20}$$

Combining (16)-(20) results in

$$\begin{aligned}
&KL\big(\mathbf{Q}(Y_0^a:Y_t^a)||\mathbf{P}(Y_0:Y_t)\big)\le\\
&\sum_{i=0}^{t}\kappa^2\|s_0\|^2\lambda^{-2t}\lambda_{max}(\Sigma_w^{-1})+L_h^2\kappa^2\|s_0\|^2\lambda^{-2t}\lambda_{max}(\Sigma_v^{-1})\\
&\le\frac{\kappa^2\|s_0\|^2}{1-\lambda^2}\big(\lambda_{max}(\Sigma_w^{-1})+L_h^2\lambda_{max}(\Sigma_v^{-1})\big)\triangleq b_\epsilon.
\end{aligned}\tag{21}$$

Finally, with b_ϵ defined as in (21) and applying Theorem 1, the attack sequence from (12) satisfies the ϵ-stealthiness condition with $\epsilon=\sqrt{1-e^{-b_\epsilon}}$. We now show that the proposed attack sequence is effective; i.e., that there exists $t'\in\mathbb{Z}_{\ge0}$ such that $\|x_{t'}^a\|\ge\alpha$ for arbitrarily large α.

To achieve this, consider the two dynamics from (10) and (13) for any $t\in\mathbb{Z}_{\ge0}$

$$\begin{aligned}
x_{t+1}^a&=f(x_t^a,u_t^a)+w_t=f_u(x_t^a,U_t^a)\\
e_{t+1}&=f(e_t,u_t^a)+w_t=f_u(e_t,U_t^a),
\end{aligned}\tag{22}$$

with $U_t^a=\begin{bmatrix}u_t^{aT} & w_t^T\end{bmatrix}^T$, for $t\in\mathbb{Z}_{\ge0}$. Since we assumed that the open-loop system (5) is IU on the set $\mathbf{S}$, it holds that for all $x_0^a=x_0\in\mathbf{S}$, there exits a nonzero s_0 such that for any $M>0$

$$\|x^a(t,x_0^a,U^a)-e(t,x_0^a-s_0,U^a)\|\ge M\tag{23}$$

holds in $t\ge t'$, for some $t'\in\mathbb{Z}_{\ge0}$. On the other hand, we showed in (15) that $\|x(t,x_0,U)-e(t,x_0^a-s_0,U^a)\|\le\kappa\|s_0\|\lambda^{-t}$. Combining this with (23) and using the fact that $\|x(t,x_0,U)\|\le R_{\mathbf{S}}$ results in

$$\begin{aligned}
&\|x^a(t,x_0^a,U^a)-x(t,x_0-s_0,U)\|=\\
&\|x^a(t,x_0^a,U^a)-e(t,x_0^a-s_0,U^a)+e(t,x_0^a-s_0,U^a)\\
&-x(t,x_0-s_0,U)\|\ge\|x^a(t,x_0^a,U^a)-e(t,x_0^a-s_0,U^a)\|\\
&-\|e(t,x_0^a-s_0,U^a)-x(t,x_0-s_0,U)\|\ge M-\kappa\|s_0\|\lambda^{-t}\\
&\Rightarrow\|x^a(t,x_0^a,U^a)\|\ge M-\kappa\|s_0\|\lambda^{-t}-R_{\mathbf{S}}\\
&\qquad\qquad\ge M-\kappa\|s_0\|-R_{\mathbf{S}}.
\end{aligned}$$

Since M is arbitrarily, we can choose it to satisfy $M>\alpha+R_s+\kappa\|s_0\|$, for arbitrarily large α. Thus, the system is (ϵ,α)-attackable. □

From (13), we can see that the false sensor measurements are generated by the evolution of e_t. Therefore, intuitively, the attacker wants to fool the system into believing that e_t is the actual state of the system instead of x_t^a. Since e_t and x_t (i.e., the system state if no attack occurs during $t\in\mathbb{Z}_{\ge0}$) converge to each other exponentially fast, the idea is that the system almost believes that x_t is the system state (under attack), while the actual state x_t^a becomes arbitrarily large.

Further, all parameters κ, λ, L_h, Σ_w, and Σ_v in (21) are some constants that depend either on system properties (L_h, Σ_w, and Σ_v) or are determined by the controller design (κ, λ). However, s_0 is set by the attacker, and *it can be chosen arbitrarily small to make ϵ arbitrarily close to zero*. Yet, s_0 can not be equal to zero; in that case (23) would not hold – i.e., the attack would not be impactful. Thus, unlike the attack methods targeting the prediction covariance in [12] where the attack impact linearly changes with ϵ, here arbitrarily large α (high impact attacks) can be achieved even with an arbitrarily small ϵ – it may only take more time to get to $\|x_{t'}^a\|\ge\alpha$.

Remark 2. *Even though we assumed that the closed-loop dynamics is IES, slightly weaker results can still be obtained for closed-loop dynamics with incrementally asymptotic stability. We will consider this case as avenue of future work.*

VI. Simulation Results

We illustrate our results on a case-study. Specifically, we consider a fixed-base inverted pendulum equipped with an EKF used to estimate the states of the system followed by a feedback full state controller to keep the pendulum rod in the inverted position. Using $x_1=\theta$ and $x_2=\dot\theta$, the inverted pendulum dynamics can be modeled as

$$\begin{aligned}
\dot x_1&=x_2\\
\dot x_2&=\frac{g}{r}\sin x_1-\frac{b}{mr^2}x_2+\frac{L}{mr^2};
\end{aligned}\tag{24}$$

here, θ is the angle of pendulum rod from the vertical axis measured clockwise, b is the Viscous friction coefficient, r

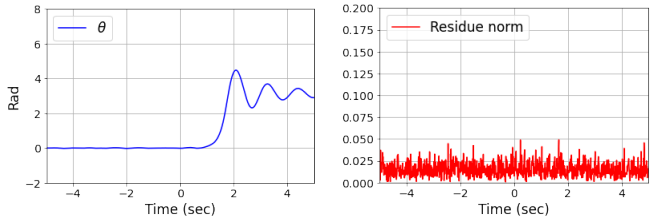

Fig. 2: (a) Angle's (θ) absolute value over time for the system under-attack, when the attack starts at time zero; (b) The residue norm over time for the system under-attack, when the attack starts at time zero.

is the radius of inertia of the pendulum about the fixed point, m is the mass of the pendulum, g is the acceleration due to gravity, and L is the external torque that is applied at the fixed base. We assumed that both the states are measured by sensors. Finally, we assumed $g = 9.8$, $m = .2Kg$, $b = .1$, $r = .3m$, $\Sigma_w = \Sigma_v = \begin{bmatrix} .01 & 0 \\ 0 & .01 \end{bmatrix}$ and discretized the model with $T_s = 10\ ms$. We assume the safe region for angle around the equilibrium point $\theta = 0$ is $\mathbf{S} = (-\frac{\pi}{3}, \frac{\pi}{3})$. To detect the presence of attack, we designed a standard χ^2-based anomaly detector that receives the sensor values and outputs the residue/anomaly alarm.

We used the attack model introduced in (12) to generate the sequence of false-data injection attacks over time. Fig. 2(a) presents the angle of the pendulum rod over time. Before the attack starts at time zero, the pendulum rod is around the angle zero; however, after initiating the attack it can be observed that the absolute value of the angle increases over time until it leaves the safe set and even becomes more than π. Note that having values more than π does not make a difference because we have a periodic system, and π corresponds to the pendulum falling down. Meanwhile, the distribution of the norm of the residue signal (see Fig. 2(b)) does not change before and after attack initiation – i.e., the attack remains stealthy.

VII. Conclusion and Future Work

In this paper, we have considered the problem of vulnerability analysis for nonlinear control systems with Gaussian noise, when attacker can compromise sensor measurements from any subset of sensors. Notions of strict stealthiness and ϵ-stealthiness have been defined, and we have shown that these notions are independent of the deployed intrusion detector. Using the KL-divergence, we have presented conditions for the existence of stealthy yet effective attacks. Specifically, we have defined the (ϵ, α)-successful attacks where the goal of the attacker is to be ϵ-stealthy while moving the system states into an unsafe region, determined by the parameter α. We have then derived a condition for which there exists a sequence of such (ϵ, α)-successful false-data injection attacks. In particular, we showed that if the closed-loop system is incrementally exponentially stable and the open-loop system is incrementally unstable, then there exists a sequence of (ϵ, α)-successful attacks.

References

[1] T. Chen and S. Abu-Nimeh, "Lessons from stuxnet," *Computer*, vol. 44, no. 4, pp. 91–93, 2011.

[2] Mo, Yilin and Sinopoli, Bruno, "False data injection attacks in control systems," in *First workshop on Secure Control Systems*, 2010, pp. 1–6.

[3] I. Jovanov and M. Pajic, "Relaxing integrity requirements for attack-resilient cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 64, no. 12, pp. 4843–4858, Dec 2019.

[4] C. Kwon, W. Liu, and I. Hwang, "Analysis and design of stealthy cyber attacks on unmanned aerial systems," *Journal of Aerospace Information Systems*, vol. 11, no. 8, pp. 525–539, 2014.

[5] A. Khazraei and M. Pajic, "Attack-resilient state estimation with intermittent data authentication," *Automatica*, vol. 138, 2022.

[6] A. Khazraei and M. Pajic, "Perfect attackability of linear dynamical systems with bounded noise," in *American Control Conf. (ACC)*, 2020.

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

[8] J. Shang and T. Chen, "Optimal stealthy integrity attacks on remote state estimation: The maximum utilization of historical data," *Automatica*, vol. 128, p. 109555, 2021.

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

[10] 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.

[11] T. Sui, Y. Mo, D. Marelli, X. Sun, and M. Fu, "The vulnerability of cyber-physical system under stealthy attacks," *IEEE Transactions on Automatic Control*, vol. 66, no. 2, pp. 637–650, 2020.

[12] C.-Z. Bai, V. Gupta, and F. Pasqualetti, "On kalman filtering with compromised sensors: Attack stealthiness and performance bounds," *IEEE Trans. on Aut. Control*, vol. 62, no. 12, pp. 6641–6648, 2017.

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

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

[15] K. Zhang, C. Keliris, T. Parisini, and M. M. Polycarpou, "Stealthy integrity attacks for a class of nonlinear cyber-physical systems," *IEEE Transactions on Automatic Control*, 2021.

[16] A. Khazraei, S. Hallyburton, Q. Gao, Y. Wang, and M. Pajic, "Learning-based vulnerability analysis of cyber-physical systems," *International Conference on Cyber-Physical Systems (ICCPS)*, 2022.

[17] A. Khazraei, H. Pfister, and M. Pajic, "Resiliency of perception-based controllers against attacks," in *Learning for Dynamics and Control Conference*. PMLR, 2022, pp. 713–725.

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

[19] D. N. Tran, B. S. Rüffer, and C. M. Kellett, "Convergence properties for discrete-time nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 64, no. 8, pp. 3415–3422, 2018.

[20] D. N. Tran, B. S. Rüffer, and C. M. Kellett, "Incremental stability properties for discrete-time systems," in *2016 IEEE 55th Conference on Decision and Control (CDC)*. IEEE, 2016, pp. 477–482.

[21] A. J. Kerns, D. P. Shepard, J. A. Bhatti, and T. E. Humphreys, "Unmanned aircraft capture and control via gps spoofing," *Journal of Field Robotics*, vol. 31, no. 4, pp. 617–636, 2014.

[22] V. Lesi, I. Jovanov, and M. Pajic, "Network scheduling for secure cyber-physical systems," in *2017 IEEE Real-Time Systems Symposium (RTSS)*, Dec 2017, pp. 45–55.

[23] A. Teixeira, D. Pérez, H. Sandberg, and K. H. Johansson, "Attack models and scenarios for networked control systems," in *First Int. Conf. on High Confidence Networked Systems*, 2012, pp. 55–64.

[24] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *47th Annual Allerton Conference on Communication, Control, and Computing*, 2009, pp. 911–918.

[25] A. Khazraei, H. Pfister, and M. Pajic, "Attacks on perception-based control systems: Modeling and fundamental limits," *arXiv preprint arXiv:2206.07150*, 2022.

[26] M. Thomas and A. T. Joy, *Elements of information theory*. Wiley-Interscience, 2006.