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 start by defining 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) are characterized by the tight integration of controllers and physical plants, potentially through communication networks. As such, they have been shown to be vulnerable to various types of cyber and physical attacks with disastrous impact (e.g., [1]). Consequently, as part of the control design and analysis process, it is critical to identify early any vulnerability of the considered system to impactful attacks, especially the ones that are potentially stealthy to the deployed intrusion detection mechanisms.

Depending on attacker capabilities, different types of stealthy attacks have been proposed. For instance, when only sensor measurements can be compromised by the attacker, it has been shown that 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, the attacker that compromises the control input can design effective stealthy attacks 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. When the attacker compromises both the plant's actuation and sensing, e.g., [11] derives the conditions under which the system is vulnerable to stealthy attacks.

However, the common assumption for all these results is that the considered plant is an LTI system. Furthermore, 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. In addition, the authors show that a sufficient condition for such notion of stealthiness is that the Kullback–Leibler (KL) divergence between the probability distribution of compromised system measurements and the attack-free measurements is close to zero, and consider stealthiness of such attacks on control systems with an LTI plant and an LQG controller.

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. Even 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. More recently, [15] introduced stealthy attacks on a specific class of nonlinear systems with residualbased intrusion detector, but provided effective attacks only when *both* plant's inputs and outputs are compromised by the attacker. 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 methods to design the stealthy attacks 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 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. To the best of our knowledge, this is the first work that considers the problem of stealthy impactful sensor attacks for systems with general nonlinear dynamics that is independent of the deployed intrusion detector. The main contributions of the paper are twofold. First, we introduce the notions of *strict* and ϵ -stealthiness. Second, using the well-known results for incremental stability introduced in [18], we derive conditions

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

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 openloop plant is incrementally unstable, then the closed-loop system is strictly vulnerable to stealthy sensing attacks.

The paper is organized as follows. In Section II, we introduce preliminaries, whereas Section III presents the system and attack model, before formalizing the notion of stealthiness in Section IV. Section V provides sufficient conditions for existence of the impactful yet stealthy attacks. Finally, in Section VI, we illustrate our results on a case-study, before concluding remarks in Section VII.

Notation: We use $\mathbb{R}, \mathbb{Z}, \mathbb{Z}_{t\geq 0}$ to 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. For a vector $x \in \mathbb{R}^n$, $||x||_p$ denotes the *p*-norm of x; 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 $||f(x) - f(y)|| \leq L||x - y||$. Finally, if **P** and **Q** are probability distributions relative to Lebesgue measure with densities **p** and **q**, respectively, then the Kullback–Leibler (KL) divergence between **P** and **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

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

where $f : \mathbb{X} \times \mathbb{D} \to \mathbb{X}$ is continuous and f(0,0) = 0. We denote by $x(t,\xi,d)$ 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 $\{d_0 : d_{t-1}\}$.¹

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 there exist $\kappa > 1$ and $\lambda > 1$ such that

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

holds for all $\xi_1, \xi_2 \in \mathbb{X}$, any $d_t \in \mathbb{D}$, and $t \in \mathbb{Z}_{t \ge 0}$. 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,

$$||x(t,\xi_1,d) - x(t,\xi_2,d)|| \ge M,$$
(3)

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

¹To simplify our notation, we denote the sequence $\{d_0 : d_{t-1}\}$ as d.



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

III. SYSTEM MODEL

In this section, we introduce the considered system and attack model, allowing us to formally capture 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 that the states of the system evolve following a general nonlinear discrete-time dynamics that can be captured in the state-space form as

$$x_{t+1} = f(x_t, u_t) + w_t, y_t = h(x_t) + v_t;$$
(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 his Lipschitz with a constant L_h . The plant output vector captures measurements from the set of plant sensors S. Furthermore, $w \in \mathbb{R}^n$ 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

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

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). Thus, it is critical to stabilize such systems using a proper controller. Due to their robustness to uncertainties, closed-loop controllers are utilized in most control systems. In

the most general form, a feedback controller design can be captured in the state-space form as

$$\begin{aligned} X_t &= f_c(X_{t-1}, y_c^c), \\ u_t &= h_c(X_t, y_t^c), \end{aligned}$$
(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 $y^c = y^2$. Note that the control model (6) is 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 control system as $\mathbf{X} \triangleq \begin{bmatrix} x \\ \chi \end{bmatrix}$, and exogenous disturbances as $\mathbf{W} \triangleq \begin{bmatrix} w \\ v \end{bmatrix}$; 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 cases, the attacker can launch false-date injection attacks, inserting a desired value instead of the current measurement of a compromised sensor.³

Thus, assuming that the attack starts at time t = 0, the sensor measurements delivered to the controller for $t \ge 0$ can be modeled as [23]

$$y_t^{c,a} = y_t^a + a_t; (8)$$

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 $\mathcal{K} = \mathcal{S}$; for some systems, we will discuss how the results can be generalized for the case when $\mathcal{K} \subset \mathcal{S}$.

Note that since the controller uses the received sensing information 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 $\mathbf{X}^{a} \triangleq \begin{bmatrix} x^{a} \\ a \end{bmatrix}$.

 $\mathbf{X} = \begin{bmatrix} \chi^a \end{bmatrix}$. 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 , $t \ge 0$, 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 system attacks (and anomalies), we assume that an intrusion detector (ID) is employed, analyzing the received sensor measurements and internal state of the controller. Specifically, by defining $Y \triangleq \begin{bmatrix} y^c \\ \chi \end{bmatrix}$, as well as $Y^a \triangleq \begin{bmatrix} y^{c,a} \\ \chi^a \end{bmatrix}$ when the system is under attack, we assume that the intrusion detector has access to a sequence of values $Y_{-\infty} : Y_t$ until time t and solves the binary hypothesis checking

 H_0 : normal condition (the ID receives $Y_{-\infty} : Y_t$); H_1 : abnormal behaviour (receives $Y_{-\infty} : Y_{-1}, Y_0^a : Y_t^a$).⁴

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 **P**, or from an **unknown** distribution of the alternative hypothesis H_1 , which we denote as **Q**. Note here that, for known noise profiles, the distribution **Q** is controlled by the injected attack signal.

Defining the intrusion detector mapping as $D : \bar{Y}^t \rightarrow \{0, 1\}$, two possible errors may occur. The error type (I) known as *false alarm*, occurs if $D(\bar{Y}^t) = 1$ when $\bar{Y}^t \sim \mathbf{P}$ and error type (II), also known as *miss-detection*, occurs when $D(\bar{Y}^t) = 0$ for $\bar{Y}^t \sim \mathbf{Q}$. Hence, we define the sum of conditional error probabilities of the intrusion detector for a given random sequence Y^t , at time t as

$$p_t^e = \mathbb{P}(D(\bar{Y}^t) = 0 | \bar{Y}^t \sim \mathbf{Q}) + \mathbb{P}(D(\bar{Y}^t) = 1 | \bar{Y}^t \sim \mathbf{P}).$$
(9)

Note that p_t^e is not a probability measure as it can take values larger than one. However, it will be useful when we define the notion of stealthy attacks in the following section.

IV. FORMALIZING STEALTHY ATTACKS REQUIREMENTS

In this section, we capture the conditions for which an attack sequence is stealthy even from an optimal intrusion detector. Specifically, we define an attack to be strictly

 $^{^{2}}$ Here we assume that the employed communication network is reliable (e.g., wired).

 $^{^3}We$ refer to sensors from ${\cal K}$ as compromised, even if a sensor itself is not directly compromised but its measurements may be altered due to e.g., network-based attacks.

⁴Since 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.

stealthy if there exists no detector that can perform better than random guess between the two hypothesis; by better we mean the true attack detection probability is higher than the false alarm probability. However, reaching such stealthiness guarantees may not be possible in general. Therefore, we define the notion of ϵ -stealthiness, which as we will show later, is attainable for a large class of nonlinear systems.

Before formally defining the notion of attack stealthiness, we introduce the following lemma.

Lemma 1. Any intrusion detector D cannot perform better than a random guess between the two hypothesis if and only if $p^e \ge 1$. Also, $p^e = 1$ if and only if D performs as well as a random guess detector.

Proof. First, we consider the case $p^e > 1$. From (9), we have

$$1 < p^{e} = \mathbb{P}(D(\bar{Y}) = 0 | \bar{Y} \sim \mathbf{Q}) + \mathbb{P}(D(\bar{Y}) = 1 | \bar{Y} \sim \mathbf{P})$$

= $1 - \mathbb{P}(D(\bar{Y}) = 1 | \bar{Y} \sim \mathbf{Q}) + \mathbb{P}(D(\bar{Y}) = 1 | \bar{Y} \sim \mathbf{P})$
(10)

Thus, $\mathbb{P}(D(\bar{Y}) = 1 | \bar{Y} \sim \mathbf{Q}) < \mathbb{P}(D(\bar{Y}) = 1 | \bar{Y} \sim \mathbf{P})$. This means the probability of attack detection is less than the false alarm rate; therefore, D is performing worse than random guess as in random guess we have $\mathbb{P}(D(\bar{Y}) = 1 | \bar{Y} \sim \mathbf{Q}) = \mathbb{P}(D(\bar{Y}) = 1 | \bar{Y} \sim \mathbf{P}) = \mathbb{P}(D(\bar{Y}) = 1)$ because random guess is independent of the given distribution. When the equality holds (i.e., $p^e = 1$), it holds that $\mathbb{P}(D(\bar{Y}) = 1 | \bar{Y} \sim \mathbf{Q}) = \mathbb{P}(D(\bar{Y}) = 1 | \bar{Y} \sim \mathbf{P})$ where the decision of the detector D is independent of the distribution of \bar{Y} and therefore, the detector performs as the random guess detector.

Since the reverse of all these implications hold, the other (i.e., necessary) conditions of the theorem also hold. \Box

Now, using Lemma 1, we can define the notions of strict stealthiness and ϵ -stealthiness as follows.

Definition 3. Consider the system from (4). An attack sequence is strictly stealthy if there exists no detector such that the total error probability p_t^e satisfies $p_t^e < 1$, for any $t \in \mathbb{Z}_{\geq 0}$. An attack is ϵ -stealthy if for a given $\epsilon > 0$, there exists no detector such that $p_t^e < 1 - \epsilon$, for any $t \in \mathbb{Z}_{>0}$.

The following theorem uses Neyman-Pearson lemma to capture the condition for which the received sensor measurements satisfy the stealthiness condition in Definition 3.

Theorem 1 ([17]). An attack sequence is

- strictly stealthy if and only if $KL(\mathbf{Q}(Y_0^a : Y_t^a)||\mathbf{P}(Y_0 : Y_t)) = 0$ 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

$$KL\left(\mathbf{Q}(Y_0^a:Y_t^a)||\mathbf{P}(Y_0:Y_t)\right) \le \log(\frac{1}{1-\epsilon^2}).$$
(11)

Remark 1. The ϵ -stealthiness condition defined in [12], [13] requires

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

This allows for the KL divergence to linearly increase over time for any $\epsilon > 0$, and as a result, 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; thus, it introduces a stronger notion of stealthiness for the attack.

A. Formalizing Attack Goal

As previously discussed, the attacker intends to maximize degradation of control performance. Specifically, 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 the unsafe region as $\mathbf{U} = \{x \in \mathbb{R}^n \mid ||x||_2 \ge \alpha\}$ for some $\alpha > R_{\mathbf{S}}$, where $R_{\mathbf{S}}$ is the radius of the safe region **S**. Moreover, the attacker wants to remain stealthy (i.e., undetected by the intrusion detector), as formalized below.

Definition 4. The attack sequence, denoted by $\{a_0, a_1, ...\}$ is referred to as (ϵ, α) -successful attack if there exists $t' \in \mathbb{Z}_{\geq 0}$ such that $||x_{t'}^a|| \geq \alpha$ 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 , the system is referred to as a perfectly attackable system.

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, \ldots resulting in $||x_t^a|| \ge \alpha$ for some $t \in \mathbb{Z}_{\ge 0}$ – i.e., we find conditions for the system to be (ϵ, α) attackable. Here, for an attack to be stealthy, we focus on the ϵ -stealthy notion; i.e., that even the best intrusion detector 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 system (7) is incrementally exponentially stable (IES) in the set **S** and the system (5) is incrementally unstable (IU) in the set **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 attack start at t = 0, let us consider the evolutions of the system with and without attacks during $t \in \mathbb{Z}_{\geq 0}$. For the system under attack, starting at time zero, the trajectory $\mathbf{X}_{0}^{a} : \mathbf{X}_{t}^{a}$ of the system and controller states is governed by

$$x_{t+1}^{a} = f(x_{t}^{a}, u_{t}^{a}) + w_{t}, \quad y_{t}^{c,a} = h(x_{t}^{a}) + v_{t} + a_{t}$$

$$x_{t}^{a} = f_{c}(X_{t-1}^{a}, y_{t}^{c,a}), \quad u_{t}^{a} = h_{c}(X_{t}^{a}, y_{t}^{c,a}).$$
(12)

On the other hand, if the system were not under attack during $t \in \mathbb{Z}_{>0}$, 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 datainjection attack occurs during $t \in \mathbb{Z}_{\geq 0}$. Since the system and measurement noises are independent of the state, we can assume that $w_t^a = w_t$ and $v_t^a = v_t$. In this case, the dynamics of the plant and controller state evolution satisfies

$$x_{t+1} = f(x_t, u_t) + w_t, \quad y_t^c = h(x_t) + v_t, \chi_t = f_c(X_{t-1}, y_t^c), \quad u_t = h_c(X_t, y_t^c),$$
(13)

which can be captured in the compact form (7), with $X_0 =$ $|x_0|$ X_0

Now, consider the sequence of attack vectors injected in the system from (12), which are constructed using the following dynamical model

$$s_{t+1} = f(x_t^a, u_t^a) - f(x_t^a - s_t, u_t^a)$$

$$a_t = h(x_t^a - s_t) - h(x_t^a),$$
(14)

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 $y_t^{c,a} = h(x_t^a) + v_t + v_t$ $a_t = h(x_t^a - s_t) + v_t$. After defining

$$e_t \stackrel{\Delta}{=} x_t^a - s_t,\tag{15}$$

and combining (14) with (12), the dynamics of e_t and the controller, and the corresponding input and output satisfy

$$e_{t+1} = f(e_t, u_t^a) + w_t, \quad y_t^{c,a} = h(e_t) + v_t, \chi_t^a = f_c(\chi_{t-1}^a, y_t^{c,a}), \quad u_t^a = h_c(\chi_t^a, y_t^{c,a}),$$
(16)

with the initial condition $e_0 = x_0^a - s_0$. Now, if we define $\mathbf{X}_t^e = \begin{bmatrix} e_t \\ \chi_t^a \end{bmatrix}$, it holds that $\mathbf{X}_{t+1}^e = F(\mathbf{X}_t^e, \mathbf{W}_t).$ (17)

with $\mathbf{X}_0^e = \begin{bmatrix} e_0 \\ \chi_0^a \end{bmatrix}$. Since we have that $x_0^a = x_0$ and $\chi_0^a = \chi_0$, it holds that $\mathbf{X}_0 - \mathbf{X}_0^e = \begin{bmatrix} s_0 \\ 0 \end{bmatrix}$. On the other hand, since both (17) and (7) share the same function and argument \mathbf{W}_t , the closed-loop system (17) is IES, and it also follows that

$$\|\mathbf{X}(t, \mathbf{X}_0, \mathbf{W}) - \mathbf{X}^e(t, \mathbf{X}_0^e, \mathbf{W})\| \le \kappa \|\mathbf{X}_0 - \mathbf{X}_0^e\|\lambda^{-t} \le \kappa \|s_0\|\lambda^{-t};$$
(18)

therefore, the trajectories of 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 $\mathbf{Z}_t = \begin{bmatrix} x_t \\ y_t^c \end{bmatrix}$ and $\mathbf{Z}_t^e = \begin{bmatrix} e_t \\ y_t^{c,a} \end{bmatrix}$, it holds that $KL(\mathbf{Q}(Y_0^a:Y_t^a)||\mathbf{P}(Y_0:Y_t))$ $\stackrel{(i)}{\leq} KL(\mathbf{Q}(\mathbf{X}_0^e:\mathbf{X}_t^e)||\mathbf{P}(\mathbf{X}_0:\mathbf{X}_t))$ $\stackrel{(ii)}{\leq} KL \big(\mathbf{Q}(\mathbf{Z}_{-\infty} : \mathbf{Z}_{-1}, \mathbf{Z}_0^e : \mathbf{Z}_t^e) || \mathbf{P}(\mathbf{Z}_{-\infty} : \mathbf{Z}_{-1}, \mathbf{Z}_0 : \mathbf{Z}_t) \big),$ (19)

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).⁵ Then, we apply the chain-rule property of KL-divergence on the right-hand side of (19) to obtain the following

$$KL(\mathbf{Q}(\mathbf{Z}_{-\infty}:\mathbf{Z}_{-1},\mathbf{Z}_{0}^{e}:\mathbf{Z}_{t}^{e})||\mathbf{P}(\mathbf{Z}_{-\infty}:\mathbf{Z}_{-1},\mathbf{Z}_{0}:\mathbf{Z}_{t}))$$

$$= KL(\mathbf{Q}(\mathbf{Z}_{-\infty}:\mathbf{Z}_{-1})||\mathbf{P}(\mathbf{Z}_{-\infty}:\mathbf{Z}_{-1}))+$$

$$KL(\mathbf{Q}(\mathbf{Z}_{0}^{e}:\mathbf{Z}_{t}^{e}|\mathbf{Z}_{-\infty}:\mathbf{Z}_{-1})||\mathbf{P}(\mathbf{Z}_{0}:\mathbf{Z}_{t}|\mathbf{Z}_{-\infty}:\mathbf{Z}_{-1})))$$

$$= KL(\mathbf{Q}(\mathbf{Z}_{0}^{e}:\mathbf{Z}_{t}^{e}|\mathbf{Z}_{-\infty}:\mathbf{Z}_{-1})||\mathbf{P}(\mathbf{Z}_{0}:\mathbf{Z}_{t}|\mathbf{Z}_{-\infty}:\mathbf{Z}_{-1}));$$
(20)

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.

Applying the chain-rule property of KL-divergence to (20) results in

$$KL\left(\mathbf{Q}(\mathbf{Z}_{0}^{e}:\mathbf{Z}_{t}^{e}|\mathbf{Z}_{-\infty}:\mathbf{Z}_{-1})||\mathbf{P}(\mathbf{Z}_{0}:\mathbf{Z}_{t}|\mathbf{Z}_{-\infty}:\mathbf{Z}_{-1})\right)$$

$$\leq KL\left(\mathbf{Q}(e_{0}|\mathbf{Z}_{-\infty}:\mathbf{Z}_{-1})||\mathbf{P}(x_{0}|\mathbf{Z}_{-\infty}:\mathbf{Z}_{-1})\right)$$

$$+ KL\left(\mathbf{Q}(y_{0}^{e,a}|e_{0},\mathbf{Z}_{-\infty}:\mathbf{Z}_{-1})||\mathbf{P}(y_{0}|x_{0},\mathbf{Z}_{-\infty}:\mathbf{Z}_{-1})\right)$$

$$+ \dots + KL\left(\mathbf{Q}(e_{t}|\mathbf{Z}_{-\infty}:\mathbf{Z}_{t-1}^{e})||\mathbf{P}(x_{t}|\mathbf{Z}_{-\infty}:\mathbf{Z}_{t-1})\right)$$

$$+ KL\left(\mathbf{Q}(y_{t}^{e,a}|e_{t},\mathbf{Z}_{-\infty}:\mathbf{Z}_{t-1}^{e})||\mathbf{P}(y_{t}|x_{t},\mathbf{Z}_{-\infty}:\mathbf{Z}_{t-1})\right).$$

$$(21)$$

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 (13) and (16), 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 (18) we showed that $||x_t - e_t|| \leq \kappa ||s_0||\lambda^{-t}$ holds for $t \in \mathbb{Z}_{>0}$. Therefore, for all $t \in \mathbb{Z}_{\geq 0}$, it holds that

$$KL(\mathbf{Q}(e_t|\mathbf{Z}_{-\infty}:\mathbf{Z}_{t-1}^e)||\mathbf{P}(x_t|\mathbf{Z}_{-\infty}:\mathbf{Z}_{t-1})) = = (x_t - e_t)^T \Sigma_w^{-1}(x_t - e_t)$$
(22)
$$\leq \kappa^2 ||s_0||^2 \lambda^{-2t} \lambda_{max}(\Sigma_w^{-1}),$$

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

Now, using the Markov property it holds that $\mathbf{Q}(y_t^{c,a}|e_t,\mathbf{Z}_{-\infty} \quad : \quad \mathbf{Z}_{t-1}^e) \quad = \quad \mathbf{Q}(y_t^{c,a}|e_t)$ and $\mathbf{P}(y_t|x_t, \mathbf{Z}_{-\infty} : \mathbf{Z}_{t-1}) = \mathbf{P}(y_t|x_t);$ also, from (13) and (16) 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

$$KL(\mathbf{Q}(y_t^{c,a}|e_t)||\mathbf{P}(y_t|x_t)) = (h(x_t) - h(e_t))^T \Sigma_v^{-1} (h(x_t) - h(e_t))$$

$$\leq L_h^2 (x_t - e_t)^T \Sigma_v^{-1} (x_t - e_t)$$

$$\leq L_h^2 \kappa^2 \|s_0\|^2 \lambda^{-2t} \lambda_{max} (\Sigma_v^{-1}).$$
(23)

⁵Due to the space limitation, we do not introduce data-processing, chainrule, and monotonicity properties of KL-divergence. More information about these terms can be found in [25].

Combining (19)-(23) results in

$$KL\left(\mathbf{Q}(Y_0^a:Y_t^a)||\mathbf{P}(Y_0:Y_t)\right) \leq \sum_{i=0}^t \kappa^2 ||s_0||^2 \lambda^{-2t} \lambda_{max}(\Sigma_w^{-1}) + L_h^2 \kappa^2 ||s_0||^2 \lambda^{-2t} \lambda_{max}(\Sigma_v^{-1}) \leq \frac{\kappa^2 ||s_0||^2}{1-\lambda^2} \left(\lambda_{max}(\Sigma_w^{-1}) + L_h^2 \lambda_{max}(\Sigma_v^{-1})\right) \stackrel{\Delta}{=} b_{\epsilon}.$$
(24)

Finally, with b_{ϵ} defined as in (24), the attack sequence defined in (14) satisfies the ϵ -stealthiness condition with $\epsilon = \sqrt{1 - e^{-b_{\epsilon}}}$.

We now show that the proposed attack sequence is effective; i.e., there exists $t' \in \mathbb{Z}_{\geq 0}$ such that $||x_{t'}^a|| \geq \alpha$ for arbitrarily large α . To achieve this, consider the two dynamics from (12) and (16) for any $t \in \mathbb{Z}_{\geq 0}$

$$x_{t+1}^{a} = f(x_{t}^{a}, u_{t}^{a}) + w_{t} = f_{u}(x_{t}^{a}, U_{t}^{a})$$

$$e_{t+1} = f(e_{t}, u_{t}^{a}) + w_{t} = f_{u}(e_{t}, U_{t}^{a})$$
(25)

with $U_t^a = \begin{bmatrix} u_t^{aT} & w_t^T \end{bmatrix}^T$, for $t \in \mathbb{Z}_{\geq 0}$. Since we assumed that the open-loop system (5) is IU on the set **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

$$\|x^{a}(t, x_{0}^{a}, U^{a}) - e(t, x_{0}^{a} - s_{0}, U^{a})\| \ge M$$
(26)

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

On the other hand, we showed in (18) that $||x(t, x_0, U) - e(t, x_0^a - s_0, U^a)|| \le \kappa ||s_0||\lambda^{-t}$. Combining this with (26) and using the fact that $||x(t, x_0, U)|| \le R_{\mathbf{S}}$ results in

$$\begin{aligned} \|x^{a}(t, x_{0}^{a}, U^{a}) - x(t, x_{0} - s_{0}, U)\| &= \\ \|x^{a}(t, x_{0}^{a}, U^{a}) - e(t, x_{0}^{a} - s_{0}, U^{a}) + e(t, x_{0}^{a} - s_{0}, U^{a}) \\ - x(t, x_{0} - s_{0}, U)\| &\geq \|x^{a}(t, x_{0}^{a}, U^{a}) - e(t, x_{0}^{a} - s_{0}, U^{a})\| \\ - \|e(t, x_{0}^{a} - s_{0}, U^{a}) - x(t, x_{0} - s_{0}, U)\| &\geq M - \kappa \|s_{0}\|\lambda^{-t} \\ &\Rightarrow \|x^{a}(t, x_{0}^{a}, U^{a})\| \geq M - \kappa \|s_{0}\|\lambda^{-t} - R_{\mathbf{S}} \\ &\geq M - \kappa \|s_{0}\| - R_{\mathbf{S}}. \end{aligned}$$

$$(27)$$

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 (16), 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 $t \in \mathbb{Z}_{\geq 0}$) 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.

Furthermore, all parameters κ , λ , L_h , Σ_w , and Σ_v in (24) are some constants that depend either on system properties $(L_h, \Sigma_w, \text{ and } \Sigma_v)$ 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 (26) would not

hold – i.e., the attack would not not be impactful. Therefore, as opposed to 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 $||x_{t'}^{\alpha}|| \geq \alpha$.

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 future work.

Remark 3. For constructing the attack sequence in (14) we assumed that the attacker has knowledge of the system's nonlinear functions f and h, as well as has access to the values of the system state. In future work, we will show how these assumptions can be relaxed for systems with general nonlinear dynamics.

Remark 4. In case that either w = 0 or v = 0 (i.e., when there is no process or measurement noise), one can still get a similar bound on the KL-divergence only as a function of the nonzero noise covariance by applying monotonicity and data-processing inequality. However, ensuring stealthiness requirement is not possible if both w = 0 and v = 0 (i.e., for the noiseless system), as the system would be completely deterministic, and thus theoretically any small perturbation to the sensor measurements could be detected.

A. Vulnerability Analysis of LTI Systems

Theorem 2 can also be applied to find the condition for the existence of (ϵ, α) -successful attacks on LTI systems. Specifically, the LTI formulation of (4) and (6) is

$$x_{t+1} = Ax_t + Bu_t + w_t, \quad y_t^c = Cx_t + v_t, \chi_t = A_c \chi_{t-1} + B_c y_t^c, \quad u_t = C_c \chi_t;$$
(28)

LTI systems with any controller (e.g., LQG controllers) can be captured in the above form. The following lemma provides the conditions for IES and IU for the above LTI system.

Lemma 2. Consider the LTI dynamical system in the form of $x_{t+1} = Ax_t + Bd_t$. The system is IES if and only if all eigenvalues of the matrix A are inside the unit circle. The system is IU if and only if A has an unstable eigenvalue.

Proof. The proof is straightforward and follows from the definition and the direct method of Lyapunov. \Box

This allows us to directly capture conditions for stealthy yet effective attacks on LTI systems.

Corollary 1. The LTI system (28) is (ϵ, α) -attackable for arbitrarily large α if the matrix A is unstable and the closed-loop control system is asymptotically stable.

Proof. The proof is directly obtained by combining Theorem 2 and Lemma 2. \Box

Asymptotic stability of the closed-loop system is not a restrictive assumption as stability is commonly the weakest required performance guarantee for a control system. Matrix A being unstable is a necessary and sufficient condition for satisfying (ϵ, α) -attackability when any set of sensors can be compromised. Note that the (ϵ, α) -attackability condition for LTI systems with an optimal detector complies with the results from [2], [3] where LQG controllers with residue based detectors (e.g., χ^2 detectors) have been considered.

Remark 5. The false-date injection attack sequence design method from (14) will reduce into a simple dynamical model

$$s_{t+1} = Ax_t^a + Bu_t^a - (A(x_t^a - s_t) + Bu_t^a) = As_t$$

$$a_t = C(x_t^a - s_t) - C(x_t^a) = -Cs_t,$$
(29)

that only requires knowledge about the matrices A and C. In addition, unlike the case for nonlinear systems, there is no need to have access to the actual states of the system.

Remark 6. In Section III-A.3 we assumed that $\mathcal{K} = S$; i.e., the attacker can compromise all sensors. However, when the system is LTI, the minimum subset of compromised sensors can be obtained as

$$\min_{v_i \in \{v_1, \dots, v_q\}} \|supp(Cv_i)\|_0,$$
(30)

where $\{v_1, ..., v_q\}$ denotes the set of unstable eigenvectors of the matrix A, and supp denotes the set of nonzero elements of the vector.

VI. SIMULATION RESULTS

We illustrate our results on a case-study. Specifically, we consider a fixed-base inverted pendulum equipped with an Extended Kalman Filter 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

$$x_1 = x_2
\dot{x}_2 = \frac{g}{r} \sin x_1 - \frac{b}{mr^2} x_2 + \frac{L}{mr^2};$$
(31)

here, θ is the angle of pendulum rod from the vertical axis measured clockwise, b is the Viscous friction coefficient, ris 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 considered in (14) to generate the sequence of false-data injection attacks over time. Fig. 2(a) shows the angle of the pendulum pod over time. Before the attack starts at time zero, the pendulum pod 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



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

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. We also provided the results for LTI systems, showing that they are compatible with the existing results for LTI systems and χ^2 based detectors.

Our results assume that the attacker has knowledge of the state evolution function f, as well as access to the values of the actual system states and the control inputs during the attack. Future work will be directed toward deriving conditions when the attacker has limited knowledge about the states, control input and the function f. We will also study the effects of specific previously reported attacks (e.g., replay attack) on general nonlinear control systems using the defined notions of strict and ϵ -stealthiness.

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 attackresilient 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*, 2021.
- [6] A. Khazraei and M. Pajic, "Perfect attackability of linear dynamical systems with bounded noise," in 2020 American Control Conference (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 Transactions on Automatic 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," Duke University, Tech. Rep., 2021, available at https://cpsl.pratt.duke.edu/publications.
- [18] D. Angeli, "A lyapunov approach to incremental stability properties," *IEEE Transactions on Automatic 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. Pajić, "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 2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton). IEEE, 2009, pp. 911–918.
- [25] M. Thomas and A. T. Joy, *Elements of information theory*. Wiley-Interscience, 2006.