Resiliency of Perception-Based Controllers Against Attacks

Amir Khazraei  
AMIR.KHAZRAEI@DUKE.EDU

Henry Pfister  
HENRY.PFISTER@DUKE.EDU

Miroslav Pajic  
MIROSLAV.PAJIC@DUKE.EDU

Department of Electrical and Computer Engineering, Duke University, Durham, NC 27705

Abstract

This work focuses on resiliency of learning-enabled perception-based controllers for nonlinear dynamical systems. We consider systems equipped with an end-to-end controller, mapping the perception (e.g., camera images) and sensor measurements to control inputs, as well as a statistical or learning-based anomaly detector (AD). We define a general notion of attack stealthiness and find conditions for which there exists a sequence of stealthy attacks on perception and sensor measurements that forces the system into unsafe operation without being detected, for any employed AD. Specifically, we show that systems with unstable physical plants and exponentially stable closed-loop dynamics are vulnerable to such stealthy attacks. Finally, we use our results on a case-study.

Keywords: resiliency of learning-enabled controllers, perception-based control, anomaly detection.

1. Introduction

Due to the recent advances in deep learning and perception, the next generation of control systems is incorporating perception modules to extract information from the environment for control and decision making. This includes end-to-end control systems that directly incorporate camera images, LiDAR 3D point clouds, and other sensor information to compute control inputs at runtime Pan et al. (2017); Rausch et al. (2017); Jaritz et al. (2018); Polvara et al. (2018); Codevilla et al. (2018), as well as controllers that first extract state information from the images followed by the use of classic feedback controllers Dean and Recht (2020); Dean et al. (2020b,a). Yet, despite the tremendous promise, resiliency of perception-based controllers to well-documented adversarial threats has not be well addressed, limiting the use of these learning-enabled controllers in real-world applications.

In particular, the main focus of adversarial machine learning has been on vulnerability of deep neural networks (DNNs) to small input perturbation, effectively focusing on robustness analysis of DNNs; e.g., targeting DNNs classification or control performance when bounded noise is added to the images in camera-based control systems. On the other hand, an attacker capable of compromising system perception/sensing would not limit their actions to bounded measurement perturbation. Moreover, little consideration has been given on potential impact of stealthy (i.e., undetectable) attacks, which are especially dangerous in the control context.

Consequently, in this work, we focus on resiliency analysis of perception-based control systems under attack. Specifically, we consider the impact stealthy false-data injection attacks can have on these learning-enabled control systems. Our notion of attack stealthiness is closely related to the one by Bai et al. (2017) for non-perception systems, where an attack is considered stealthy if and only if it is undetected by an optimal detector. However, unlike Bai et al. (2017) we do not restrict our analysis (and stealthiness requirement) only on steady-state behaviors. For different threat models
(mainly the level of information the attacker has about the system), we derive conditions that there exists a stealthy and effective attack sequence forcing the system far from the operating point. In particular, we assume the attacker knows the open-loop plant dynamics, and differentiate the cases where they have (or do not have) access to an estimation of the plant’s state during the attack. We show that under these conditions, the probability of attack remaining stealthy can be chosen arbitrarily close to one if the attacker’s state estimation error can be arbitrarily close to zero. When the attacker does not have access to an estimate of the state, the stealthiness level of the attack depends on the system’s performance in attack-free operation. Specifically, we show that unlike for linear time-invariant (LTI) systems where stealthy and effective attacks are independent of the control design, for nonlinear systems the level of stealthiness is closely related to the level of closed-loop system stability – i.e., if the closed-loop systems is more stable, the attack can have stronger stealthiness guarantees. On the other hand, the attack impact (i.e., control degradation) fully depends on the level of open-loop system instability (e.g., the size of unstable eigenvalues for LTI systems).

**Related Work.** The initial work Szegedy et al. (2013) on adversarial example generation showed that DNNs are vulnerable to small input perturbations. Afterward, the majority of works have applied this idea to adversarial attacks on physical world such as malicious stickers on traffic signs to fool the detectors and/or classifiers Eykholt et al. (2018); Song et al. (2018); Papernot et al. (2017); Sun et al. (2020). However, all these methods only consider classification tasks in a static manner; i.e., without consideration of the longitudinal (i.e., over time) system behaviours.

Recent works by Boloor et al. (2020, 2019); Sato et al. (2020); Jia et al. (2020); Yoon et al. (2021); Hallyburton et al. (2022) have studied the vulnerabilities of perception-based autonomous vehicles in longitudinal way. For instance, Boloor et al. (2020, 2019) consider autonomous vehicles with end-to-end DNN controllers that directly map perceptual inputs into the vehicle steering angle, and target the systems by painting black lines on the road. However, all these works consider specific applications and address attack impact in an ad-hoc manner, limiting the use of their results in other systems/domains. Further, they lack any consideration of attack stealthiness, as injecting e.g., adversarial patches that only maximises the disruptive impact on the control can be detected by most ADs. For instance, Cai and Koutsoukos (2020) introduce an AD that easily detects the adversarial attacks from Boloor et al. (2020). On the other hand, in this work, we focus on nonlinear system dynamics, define general notions of attack stealthiness, and introduce sufficient conditions for a perception-based control system to *not* be resilient against perception and sensing attacks.

Finally, for non-perception control systems, stealthy attacks have been well-defined in e.g., Mo and Sinopoli (2009); Mo, Yilin and Sinopoli, Bruno (2010); Teixeira et al. (2012); Khazraei and Pajic (2021); Smith (2015); Bai et al. (2017); Jovanov and Pajic (2019); Sui et al. (2020); Khazraei et al. (2021); Khazraei and Pajic (2020). However, all these work also only focus on LTI systems and linear controllers, as well as on specific AD design (e.g., $\chi^2$ detector). Recently Khazraei et al. (2021) have introduced a learning-based attack design for systems with nonlinear dynamics; yet, their work only considers stealthiness with respect to the $\chi^2$-based AD, and does not consider perception-based controllers.

**Notation.** $P$ denotes the probability for a random variable. For a square matrix $A$, $\lambda_{\text{max}}(A)$ is 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 $f : \mathbb{R}^n \rightarrow \mathbb{R}^p$ is locally Lipschitz with constant $L$ if for any $x, y \in D \subseteq \mathbb{R}^n$ it holds that $\|f(x) - f(y)\| \leq L\|x - y\|$; it is globally Lipschitz with constant $L$ if $D = \mathbb{R}^n$. For a set $X,
Resiliency of Perception-Based Controllers Against Attacks

\( \partial X \) and \( X^o \) define the boundary and the interior of the set, respectively. \( B_r \) denotes a closed ball with radius \( r \); i.e., \( B_r = \{ x \in \mathbb{R}^n | \| x \| \leq r \} \), whereas \( 1_A \) is the indicator function on a set \( A \). For a function \( f \), we denote \( f' = \frac{\partial f}{\partial x} \) as the partial derivative of \( f \) with respect to \( x \) and \( \nabla f_i(x) \) is the gradient of the function \( f_i \) (\( i \)-th element of the function \( f \)). Finally, if \( P \) and \( Q \) are probability distributions relative to the same Lebesgue measure, then the total variation between them is defined as \( \| P - Q \|_{tv} = \frac{1}{2} \int |P(x) - Q(x)| \, dx \). The Kullback–Leibler (KL) divergence between \( P \) and \( Q \) is \( KL(P, Q) = \int P(x) \log \frac{P(x)}{Q(x)} \, dx \).

2. Problem Description: System and Attack Models

In this section, we introduce the system and attack model. Specifically, we consider the setup from Fig. 1 where each of the components is modeled as follows.

2.1. Plant and Perception Model

We assume the plant can be modeled with a nonlinear dynamics in the standard state-space form

\[
x_{t+1} = f(x_t) + B u_t + w_t, \quad y_t^s = C s x_t + v_t^s, \quad z_t = G(x_t).
\]

(1)

Here, \( x_t \in \mathbb{R}^n \), \( u_t \in \mathbb{R}^m \), \( w_t \in \mathbb{R}^n \), \( z_t \in \mathbb{R}^l \), \( y_t^s \in \mathbb{R}^s \) and \( v \in \mathbb{R}^s \) denote the state, input, system disturbance, observations from perception-based sensors, (non-perception) sensor measurements, and sensor noise, at time \( t \), respectively. We assume \( f \) is globally Lipschitz with constant \( L \) and without loss of generality that \( f(0) = 0 \). The perception-based sensing is modeled by an unknown generative model \( G \), which is nonlinear and potentially high-dimensional. Finally, the process and measurement noise vectors \( w \) and \( v^s \) are independent and identically distributed (iid) Gaussian processes with \( w \sim \mathcal{N}(0, \Sigma_w) \) and \( v^s \sim \mathcal{N}(0, \Sigma_v^s) \).

For example, consider a camera-based lane keeping. Here, the observations \( z_t \) are the captured images; the map \( G \) generates the images based on the vehicle’s position and velocity.

2.2. Control Unit

The control unit, shown in Fig. 1, consists of perception, controller and anomaly detector units.

**Perception.** We assume that there exists a perception map \( P \) that imperfectly estimates the partial state information; i.e.,

\[
y_t^P = P(z_t) = C_P x_t + v^P(x_t),
\]

(2)

where \( P \) is a deep neural network (DNN) trained using any supervised learning method on a data set \( \mathcal{X} = \{(z_i, x_i)\}_{i=1}^N \) collected densely around the operating point \( x_o \) of the system, as in Dijk and Croon (2019); Lambert et al. (2018). In addition, \( v^P \in \mathbb{R}^p \) is the perception map error that may depend on the state of the system – i.e., smaller around the training data set. To capture perception guarantees, we employ the model for robust perception-based control from Dean and Recht (2020).
Specifically, if the model is trained effectively, the perception error $e$ around the operating point $x_o$ should be bounded, i.e., the following assumption from Dean et al. (2020a) holds.

**Assumption 1** There exists a safe set $S$ around the operating point such that for all $x \in S$, it holds that $\|P(z) - CP_x\| \leq \gamma_e$, where $z = G(x)$ - i.e., for all $x \in S$, $\|v^p(x)\| \leq \gamma_e$. Without loss of generality, in this work we consider the origin as the operating point - i.e., $x_o = 0$.

**Controller.** The system (1) is controlled by a (general) nonlinear controller $u_t = \pi(z_t, y_t^c)$. Hence, for $h(x_t, v_t) = f(x_t) + B\pi(z_t, y_t^c)$, the evolution of the closed loop system can be captured as

$$x_{t+1} = h(x_t, v_t) + w_t.$$  \hspace{1cm} (3)

In the general form, the controller can employ any end-end control policy that uses the image and sensor measurements. When the system is noiseless, the the state dynamics can be captured as

$$x_{t+1} = h(x_t, 0).$$  \hspace{1cm} (4)

We now introduce the following definitions that describe the properties of the controlled system.

**Definition 1** The origin of the system (4) is exponentially stable on a set $D \subseteq \mathbb{R}^n$ if for any $x_0 \in D$, there exist $0 < \alpha < 1$ and $M > 0$, such that $\|x_t\| \leq M\alpha^t\|x_0\|$, for all $t \geq 0$.

**Lemma 2** Kushner (2014) For the system of (4), if there exists a function $V : \mathbb{R}^n \to \mathbb{R}$ such that for any $x_t \in D \subseteq \mathbb{R}^n$ we have

$$c_1\|x_t\|^2 \leq V(x_t) \leq c_2\|x_t\|^2, \quad V(x_{t+1}) - V(x_t) \leq -c_3\|x_t\|^2, \quad \frac{\partial V(x)}{\partial x} \leq c_4\|x\|.$$  \hspace{1cm} (5)

for some positive $c_1$, $c_2$, $c_3$ and $c_4$, then the origin is exponentially stable.

**Assumption 2** We assume that for the closed-loop control system (4) is exponentially stable on a set $D = B_d$. Using the converse Lyapunov theorem Khalil (2002), there exists a Lyapunov function that satisfies the inequalities in (5) with constants $c_1$, $c_2$, $c_3$ and $c_4$ on a set $D = B_d$.

**Remark 3** The assumptions made for closed-loop system are critical for system guarantees without the attack; i.e., if the system does not satisfy the stability property in attack-free condition, then the best strategy for the attacker would be to wait until the system fails by itself. We refer the reader to the recent work e.g., Dean and Recht (2020); Dean et al. (2020a) on design of such controllers.

**Remark 4** Note that the exponential stability assumption for the closed-loop control system (4) can be relaxed to control systems with a quadratic-type Lyapunov function. For such control systems instead of satisfying (5), it is sufficient to have a positive definite $V(x)$ that satisfies $V(x_{t+1}) - V(x_t) \leq -c_3\phi^2(x)$ and $\|\frac{\partial V}{\partial x}\| \leq c_4\phi^2(x)$, where $\phi : \mathbb{R}^n \to \mathbb{R}$ is a positive definite and continuous function, and $c_3$, $c_4$ are positive scalars. However, to simplify our presentation, in this work we assume exponential stability of the system, although our results can be easily extended to this case.

**Definition 5** The class of functions $U_\rho$ contains all functions $f$ such that the dynamics $x_{t+1} = f(x_t) + d_t$, where $d_t$ satisfies $\|d_t\| \leq \rho$, becomes arbitrarily large for some nonzero initial state $x_0$. Also, for a function $f$ from $U_\rho$ and initial condition $x_0$, we define $T_f(\alpha, x_0) = \min\{t \mid \|x_t\| \geq \alpha\}$.

In other words, $T_f(\alpha, x_0)^1$ is the minimum time needed for an unstable dynamic $f$, subject to the initial condition $x_0$, to leave a bounded ball with radius $\alpha$ subject to the initial condition $x_0$. In the following, we elaborate the main assumptions that we make on the close-loop control system.

---

1. To simplify our notation, and since we consider specific $f$ from the plant dynamics (1), we drop the the subscript $f$. 

---

Resiliency of Perception-Based Controllers Against Attacks
**Anomaly Detector.** The system is equipped with an anomaly detector (AD) designed to detect the presence of any abnormal behaviours. We use $y_t = \begin{bmatrix} y_t^p \\ y_t^a \end{bmatrix}$ and $y_t^a = \begin{bmatrix} y_t^{P,a} \\ y_t^{S,a} \end{bmatrix}$ to capture sensor and perception-based (from (2)) values without and under attack, respectively (the full attack model is introduced in the next subsection). Now, consider the classical binary hypothesis testing problem:

- $H_0$: normal condition (the AD receives $y_0 : y_t$);
- $H_1$: abnormal behaviour (the AD receives $y_0^a : y_t^a$).

Effectively, the AD uses both the extracted state information from the perception map (i.e., (2)) and sensor measurements. Given a random sequence $Y = (y_0 : y_t)$, it either comes from the distribution $P$ (null hypothesis $H_0$) or from a distribution $Q$ (the alternative hypothesis $H_1$). For a given AD specified by function $D : Y \to \{0, 1\}$, two types of error may occur. Error type (I), also referred as false alarm, occurs if $D(Y) = 1$ when $Y \sim P$, whereas type (II) error (miss detection) occurs if $D(Y) = 0$ when $Y \sim Q$. Hence, the total error probability of AD $D$ for a given random sequence $Y$ is

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

We define the attack to be stealthy if there exists no detector that can do better than ignoring the received measurements and making a random guess to decide between the two hypotheses. Since in random guess the output of the detector is independent of whether $Y \sim P$ or $Y \sim Q$, we have $p_t^e = \mathbb{P}(D(Y) = 0) + \mathbb{P}(D(Y) = 1) = 1$. This is also equivalent to the case if there exists no detector that satisfies $p_t^e < 1$ or $\mathbb{P}(D(Y) = 1 | Y \sim Q) \geq \mathbb{P}(D(Y) = 1 | Y \sim P)$ which means the probability of true detection be greater than false alarm.

### 2.3. Attack Model

We assume that the attacker has the ability to compromise outputs of the perception module(s) (e.g., camera images) as well as (potentially) the sensor measurements $y_t^a$ (see Fig. 1). Moreover, the attack starts at $t = 0$, and we use the superscript $a$ to differentiate all signals of the attacked system, for all $t \geq 0$; the attack sequence is \{z_t^a, y_t^{S,a}\}$, where e.g., the value of observation delivered to the perception unit at time $t$ is denoted by $z_t^a$. Therefore, in the presence of an attack, the system dynamics can be captured as

$$x_{t+1}^a = f(x_t^a) + Bu_t^a + w_t^a,$$

$$u_t^a = \pi(z_t^a, y_t^{S,a}).$$

In this work, we assume the attacker has full knowledge of the system, its dynamics and employed architecture. Further, the attacker has the required computation power to calculate suitable attack signals to inject, planning ahead as needed. We formally define the attack stealthiness as follows.

**Remark 6** In our notation, $x_0^a : x_t^a$ denotes a state trajectory of the system under attack (for an attack starting at $t = 0$), while $x_0 : x_t$ denotes the state trajectory of the attack-free system; we refer to such state trajectory as the attack-free trajectory. Thus, when comparing the attack-free trajectory and the system trajectory under attack (i.e., from (7)), we assume $u_t^a = u_t$ and $v_t^{S,a} = v_t^a$.

**Definition 7** Consider the system defined in (1). 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 $t \geq 0$. An attack is **$\epsilon$-stealthy** if for a given $\epsilon > 0$, there exists no detector such that $p_t^e < 1 - \epsilon$, for any $t \geq 0$. 

5
**Theorem 8** An attack sequence is strictly stealthy if and only if $KL(Q(y_0^t : y_t^t)||P(y_0 : y_t)) = 0$ for all $t \geq 0$, where $KL$ represents the Kullback–Leibler divergence operator. An attack sequence is $\epsilon$-stealthy if the corresponding observation sequence $y_0 : y_t$ satisfies
\[
KL(Q(y_0^t : y_t^t)||P(y_0 : y_t)) \leq \log\left(\frac{1}{1 - \epsilon^2}\right).
\] (8)

**Proof** First we prove the strictly stealthy case. Using Neyman-Pearson Lemma for any existing detector $D$, it follows that
\[
p_t^e \geq \int \min\{P(y), Q(y)\} dy,
\] (9)
where the equality holds for the Likelihood Ratio function as $D^* = 1_{P \geq Q}$ Krishnamurthy (2017). Since $1 - \int \min\{P(y), Q(y)\} dy = \frac{1}{2} \int |P(x) - Q(x)| dx$, from Polianskiy and Wu (2014), from the definition of total variation distance between $P$ and $Q$, we have
\[
p_t^e \geq 1 - \|P - Q\|_{tv}.
\] (10)

Now, it holds that $\|P - Q\|_{tv} \leq \sqrt{1 - e^{-KL(Q||P)}}$ (Eq. (14.11) in Lattimore and Szepesvári (2020)). Thus, if we have $KL(P(y_0^t : y_t^t)||Q(y_0 : y_t)) = 0$, then $p_t^e \geq 1$ for any detector $D$. On the other hand, if for all detectors we have $p_t^e \geq 0$, then the equality holds for $\|P - Q\|_{tv} = 0$, which is equivalent to $P = Q$ and therefore, $KL(P(y_0^t : y_t^t)||Q(y_0 : y_t)) = 0$.

For the $\epsilon$-stealthy case, we combine (10) with the inequality $\|P - Q\|_{tv} \leq \sqrt{1 - e^{-KL(Q||P)}}$ and the $\epsilon$-stealthy condition (8), to show $p_t^e \geq 1 - \|P - Q\|_{tv} \geq 1 - \sqrt{1 - e^{-KL(Q||P)}} \geq 1 - \epsilon$. ■

**Attack Goal** is to maximize degradation of control performance. Specifically, as we consider the origin as the operating point, the attack objective is to maximize the (norm of) states $x_t$. Moreover, the attacker wants to remain stealthy – i.e., undetected by the ID, as formalized below.

**Definition 9** The attack sequence, denoted as $\{z_0^0, y_0^{s,a}\}, \{z_1^a, y_1^{s,a}\}, ...$, is referred to $(\epsilon, \alpha)$-successful attack if there exists $t' \geq 0$ such that $\|x_{t'}\| \geq \alpha$ and the attack is $\epsilon$-stealthy for all $t \geq 0$. When such a sequence exists for a system, the system is called $(\epsilon, \alpha)$-attackable. Finally, when the system is $(\epsilon, \alpha)$-attackable for arbitrarily large $\alpha$, the system is referred to as perfectly attackable.

Our goal is the derive methods to capture impact of stealthy attacks; specifically, in the next section, we derive conditions for existence of a stealthy yet effective attack sequence $\{z_0^0, y_0^{s,a}\}, \{z_1^a, y_1^{s,a}\}, ...$, resulting in $\|x_{t}\| \geq \alpha$ for some $t \geq 0$ – i.e., we find conditions for a system to be $(\epsilon, \alpha)$-attackable. Here, for an attack to be stealthy, we focus on the $\epsilon$-stealthy notion; i.e., that the best AD could only improve the probability detection by $\epsilon$ compared to random-guess baseline detector.

### 3. Conditions for $(\epsilon, \alpha)$-Attackable Systems

To provide sufficient conditions for a system to be $(\epsilon, \alpha)$-attackable, in this section, we introduce two methodologies for design of attack sequences on perception and (classical) sensing data. The difference in these strategies is the level of knowledge the attacker has about the system; we show that the stronger attack impact can be achieved with the attacker having full knowledge of the system. Specifically, we start with the attack strategy where the attacker has access to the current plant state; in such case, we show that the stealthiness condition is less restrictive, simplifying design of $\epsilon$-stealthy attacks. For the second attack strategy, we show that the attacker can launch the attack sequence with only knowing the function $f$ (i.e., plant model); however, achieving $\epsilon$-stealthy attack in this case is harder as more restrictive conditions are imposed on the attacker.
3.1. Attack Strategy I: Using Estimate of the Plant State

Consider the attack sequence where \( z_t^a \) and \( y_t^{s,a} \) injected at time \( t \), for all \( t \geq 0 \), satisfy

\[
z_t^a = G(x_t^a - s_t), \quad y_t^{s,a} = C_s(x_t^a - s_t) + v_t^{s,a}, \tag{11}
\]

with \( s_{t+1} = f(\hat{x}_t^a) - f(x_t^a) \), and \( f(\hat{x}_t^a) \) is a state estimation (in the presence of attacks), and thus \( \zeta_t = \hat{x}_t^a - x_t^a \) is the corresponding state estimation error. Note that the attacker can obtain \( \hat{x}_t^a \) by e.g., running a local estimator. We assume that the estimation error is bounded by \( b_\zeta \) i.e., \( \|\zeta_t\| \leq b_\zeta \), for all \( t \geq 0 \). On the other hand, the above attack design may not require access to the true plant state \( x_t^a \), since only the “shifted” (i.e., \( x_t^a - s_t \)) outputs of the real sensing/perception are injected. For instance, in the lane centring control (i.e., keeping the vehicle between the lanes), \( G(x_t - s_t) \) only shifts the actual image \( s_t \) to the right or left depending on the coordinate definition.

The idea behind the above attacks is to have the system believe that its (plant) state is equal to the state \( e_t \triangleq x_t^a - s_t \); thus, referred to as the fake state. Note that effectively both \( z_t^a \) and \( y_t^{s,a} \) used by an AD are directly function of the fake state \( e_t \). Thus, if the distribution of \( e_0 : e_t \) is close to \( x_0 : x_t \) (i.e., attack-free trajectory), then the attacker will be successful in designing stealthy attacks.

**Definition 10** For an attack-free state trajectory \( x_0 : x_t \), and for any \( T \geq 0 \) and \( b_x > 0 \), \( \delta(T, b_x) \) is the probability that the system state remains in the ball with radius \( b_x \) during \( 0 < t \leq T \), if it is in the ball at time \( 0 \) i.e., \( \delta(T, b_x) \triangleq \mathbb{P}(\sup_{0 < t \leq T} \|x_t\| \leq b_x \mid \|x_0\| \leq b_x) \).

The following result captures conditions under which the perception-based control system is not resilient to attacks, in the sense that it is \((\epsilon, \alpha)\)-attackable.

**Theorem 11** Consider the system (1) with closed-loop control as in Assumption 2. Assume that the functions \( f \) and \( f' \) (derivative of \( f \)) are Lipschitz and locally Lipschitz in the set \( \mathcal{D} \), with constants \( L_f \) and \( L'_f \), respectively, and let us define \( L_1 = L'_f(b_x + 2b_\zeta + d) \), \( L_2 = \min\{2L_f, L'_f(\alpha + b_x + b_\zeta + d)\} \) and \( L_3 = L'_f(b_x + d) \). Moreover, assume that \( b_x \) has the maximum value such that the inequalities \( L_1 + L_3\|B\| < \frac{c_3}{c_4} \) and \( L_2b_\zeta < \frac{c_3 - (L_1 + L_3\|B\|)c_4}{c_4} \sqrt{\frac{c_2}{c_1}} \theta r \), for some \( 0 < \theta < 1 \), are satisfied. Then, the system (1) is \((\epsilon, \alpha)\)-attackable with probability \( \delta(T, \alpha + b + b_x, s_0) \) for some \( \epsilon > 0 \), if \( f \in \mathcal{U}_\rho \) with \( \rho = 2L_f(b_x + b) \) and \( b = \frac{c_4}{c_3 - (L_1 + L_3\|B\|)c_4} \sqrt{\frac{c_2}{c_1}} \frac{L_2b_\zeta}{\theta} \).

Before proving the above theorem we, introduce the following lemma from Khalil (2002).

**Lemma 12** Khalil (2002) Let \( x = 0 \) be an exponentially stable equilibrium point of the nominal system (4). Let \( V(x_t) \) be a Lyapunov function of the nominal system that satisfies (5) in \( \mathcal{D} \), where \( \mathcal{D} = B_d \). Suppose the system is affected by additive perturbation term \( g(x_t) \) that satisfies \( \|g(x_t)\| \leq \delta + \gamma \|x_t\| \). If it holds that \( c_3 - \gamma c_4 > 0 \) with \( \delta < \frac{c_3 - \gamma c_4}{c_4} \sqrt{\frac{c_2}{c_1}} \theta d \) for all \( x \in \mathcal{D} \) and some positive \( \theta < 1 \). Then, for all \( \|x_{t_0}\| < \sqrt{\frac{c_2}{c_1}} d \), there exists \( t_1 > t_0 \), such that for all \( 0 \leq t \leq t_1 \) it holds that \( \|x_t\| \leq \kappa e^{-\beta t} \|x_0\| \) with \( \kappa = \sqrt{\frac{c_2}{c_1}} \#, \beta = \frac{(1-\theta c_3)}{2c_2} \), while for all \( t \geq t_1 \) it holds that \( \|x_t\| \leq b \) with \( b = \frac{c_4}{c_3 - \gamma c_4} \sqrt{\frac{c_2}{c_1}} \theta d \).

**Proof of Theorem 11** To show that the sequence of perturbed image and sensor values \( \{z_0^a, y_0^{s,a}\}, \{z_1^a, y_1^{s,a}\}, \ldots \) obtained by Attack Strategy II is \((\epsilon, \alpha)\)-successful attack for arbitrarily large \( \alpha \).
defining
\[ e_t = x_t^a - s_t, \]
\[ r_t = e_t - x_t, \]
we get \( z_t^0 = G(e_t), y_t^{s,a} = C_s e_t + e_t^{s,a} \) and
\[
e_{t+1} = f(x_t^0) - f(x_t^0 + \zeta_t) + f(e_t + \zeta_t) + B\pi(G(e_t), y_t^{s,a}) + w_t^0,
\]
\[
r_{t+1} = h(r_t) + f(x_t^0) - f(x_t^0 + \zeta_t) + f(e_t + \zeta_t) - f(x_t) - f(r_t)
\]
\[
+ B\pi(G(e_t), y_t^{s,a}) - B\pi(G(x_t), y_t^e) - B\pi(G(r_t), C_s r_t)
\]
\[ = h(r_t) + \sigma_1 + \sigma_2',
\]
with \( \sigma_1' = f(x_t^0) - f(x_t^0 + \zeta_t) + f(e_t + \zeta_t) - f(x_t) - f(r_t) \) and \( \sigma_2' = B\pi(G(e_t), y_t^{s,a}) - B\pi(G(x_t), y_t^{s,a}) - B\pi(G(r_t), C_s r_t) \).

Using Mean Value theorem and equality \( x_t^0 = r_t + s_t + x_t \) using equation (12), it holds that
\[
f(x_t^0 + \zeta_t) = f(x_t^0) + \partial f \big|_{(x_t^0 + \zeta_t)} \zeta_t,
\]
\[
f(e_t + \zeta_t) = f(x_t) + \partial f \big|_{(x_t)} \zeta_t,
\]
\[
f(r_t) = f(0) + \partial f \big|_{(0)} r_t = \partial f \big|_{(0)} r_t,
\]
where for instance, \( \partial f \big|_{(0)} r_t \) =
\[
\nabla f_1 (c_1 x + (1 - c_1) y) 
\]
\[
\vdots
\]
\[
\nabla f_n (c_n x + (1 - c_n) y)
\]
for some \( c_1, ..., c_n \in [0, 1] \). Thus, we get
\[
\sigma_1' = \left( \partial f \big|_{(x_t^0 + \zeta_t)} \zeta_t - \partial f \big|_{(x_t)} \zeta_t \right) + \left( \partial f \big|_{(x_t)} - \partial f \big|_{(0)} \right) r_t.
\]

Similarly, using Mean Value theorem and \( e_t = r_t + x_t \), it holds that
\[
\pi(G(e_t), y_t^{s,a}) = \pi(G(x_t), y_t^e) + \partial \pi \big|_{(x_t)} r_t,
\]
\[
\pi(G(r_t), C_s r_t) = \partial \pi \big|_{(0)} r_t.
\]

Therefore, we get
\[
\sigma_2' = B\pi \partial \pi \big|_{(x_t + \zeta_t)} - \partial \pi \big|_{(0)} r_t
\]
\[
\text{(17)}
\]

Since the functions \( f \) and \( \partial f \big|_{(x)} \) are locally Lipschitz, for all \( 0 \leq t \leq T(\alpha + b + b_x, s_0) \) we have
\[
\| \partial f \big|_{(x_t^0 + \zeta_t)} - \partial f \big|_{(x_t)} \| \leq L_f^f (\Theta(x_t^0 + \zeta_t) - \Theta(x_t) + \zeta_t) \leq L_f^f (\| x_t^0 \| + \| x_t \| + \| r_t \| + \| \zeta_t \|) \leq L_f^f (\alpha + b_x + b_x) + L_f^f \| r_t \|,\]
\[
\text{(18)}
\]

where \( \| x_t \| \leq b_x \) holds with probability \( \delta(T(\alpha + b + b_x, s_0), b_x) \). Moreover, locally Lipschitz assumption on \( f \) and the boundedness of \( \zeta \) also results in
\[
\| \partial f \big|_{(x_t^0 + \zeta_t)} - \partial f \big|_{(x_t)} \| \leq 2 L_f.
\]
\[
\text{(19)}
\]
Therefore, we get $\|\frac{\partial f}{\partial x}|_{x=x+\alpha} - \frac{\partial f}{\partial x}|_{x=x}\| \leq \min\{2L_f, L'_f(\alpha + b_x + b_\zeta) + L'_f\|r_t\|\}$ similarly, we have

\[
\|\frac{\partial f}{\partial x}|_{x=x+\alpha} - \frac{\partial f}{\partial x}|_{x=0}\| \leq L'_f(\Theta(x+r+\zeta) - \Theta(0)) \leq L'_f(\|x_t\| + \|r_t\| + \|\zeta\|)
\]

\[
\leq L'_f(b_x + b_\zeta) + L'_f\|r_t\|
\]

Thus, $\|\sigma'_t\| \leq \min\{2L_f, L'_f(\alpha + b_x + b_\zeta)\}b_\zeta + L'_f(b_x + b_\zeta)\|r_t\| + L'_f(b_x + b_\zeta)\|r_t\| + L'_f\|r_t\|^2. $

On the other hand, for all $r_t \in D$, it holds that $\|r_t\|^2 \leq d\|r_t\|$, which results in

\[
\|\sigma'_t\| \leq \min\{2L_f, L'_f(\alpha + b_x + b_\zeta)\}b_\zeta + L'_f(b_x + b_\zeta)\|r_t\| + L'_f(b_x + b_\zeta+ d)\|r_t\| = L_1\|r_t\| + L_2b_\zeta
\]

In addition, since the function $\pi'$ is locally Lipschitz, for all $x \in D$ with probability $\delta(T(\alpha + b + b_x, s_0), b_\zeta)$, it holds that $\|\frac{\partial f}{\partial x}|_{x=x_t} - \frac{\partial f}{\partial x}|_{x=0}\| \leq L'_f(\|x_t + r_t\|) \leq L'_f(\|d + b_x\|$ and we get $\|\sigma_2\| \leq L_3\|B\|\|r_t\| with $L_3 = L'_f(b_x + d); this results in $\|\sigma'_t + \sigma_2\| \leq L_2b_\zeta + (L_1 + L_3\|B\|)\|r_t\|.$

Since we have $L_1 + L_3\|B\| < \frac{c_1}{c_2}$ and $L_2b_\zeta < \frac{c_3-c(L_1 + L_3\|B\|)c_4}{2c_2}\sqrt{\frac{c_1}{c_2}}\theta r_t$, using Lemma 12 for all $\|r_0\| = \|s_0\| < \sqrt{\frac{c_1}{c_2}} d$ there exists $t_1 > 0$, such that for all $t < t_1$ we have $\|r_t\| \leq \sqrt{\frac{c_1}{c_2}} e^{-\beta t}\|s_0\|$ with $\beta = \frac{(1-\theta)(c_3-c(L_1 + L_3\|B\|)c_4)}{2c_2} \sqrt{\frac{c_1}{c_2}} L_2b_\zeta$ for $t \geq t_1$.

Now, we need to show that for $t \geq T(\alpha + b + b_x, s_0)$, it holds that $\|x_t\|^2 \geq \alpha$. Since the function $f$ is differentiable, using Mean-value theorem we get

\[
s_{t+1} = f(s_t + e_t) - f(e_t) = f(s_t) + \frac{\partial f}{\partial x}|_{s_t} e_t - f(e_t)
\]

Since $e_t = r_t + x_t$, for $0 \leq t \leq T(\alpha + b + b_x, s_0)$, we have $\|\frac{\partial f}{\partial x}|_{s_t} e_t - f(e_t)\| \leq 2L_f\|r_t + x_t\| = 2L_f(b_x + b).$ From the assumption that $f \in L_4$ with $\rho = 2L_f(b_x + b)$, there exists $s_0$ such that the $s_t$ diverges to infinity. Using Definition 5, $T(\alpha + b + b_x, s_0)$ is defined as the first (deterministic) time step where $\|s_t\| \geq \alpha + b_x + b$. On the other hand, using $e_t = x_t - s_t$ and $e_t = x_t + r_t$ we get $\|x_t\|^2 \geq \|s_t\| - \|e_t\| \geq \|s_t\| - \|x_t\| - \|r_t\| \geq \alpha$.

Now, we need to show that the designed $y_t$ satisfies the $\epsilon$-stealthiness condition. In other words, we need to show that $KL(y_t^0 : y_t^0 || y_0 : y_t) \leq \log(\frac{1}{1-e^-})$ for some $\epsilon > 0$. Since the sequences $y_0^0, ... , y_0^0$ and $y_0, ..., y_t$ are generated by $e_0, ..., e_t$ and $x_0, ..., x_t$, respectively, using the Data-processing inequality of divergence, it holds that

\[
KL(Q(y_0^0 : y_t^0 || P(y_0 : y_t))) \leq KL(Q(e_0 : e_t || P(x_0 : x_t)))
\]

On the other hand, using monotonicity property of KL-divergence we get

\[
KL(Q(e_0 : e_t || P(x_0 : x_t))) \leq KL(Q(e_0, y_0^s, a, e_1, y_1^s, a, ..., || P(x_0, y_0^s, x_1, y_1^s, ...)) || P(x_0, y_0^s, x_1, y_1^s, ...)).
\]

Now, from the Markov property and the chain rule for KL divergence, it holds that

\[
KL(Q(e_0, y_0^s, a, e_1, y_1^s, a, ..., e_t, y_t^s || P(x_0, y_0^s, x_1, y_1^s, ..., x_t, y_t^s))) \leq

KL(Q(e_0 || P(x_0)) + KL(Q(y_0^s || P(y_0^s || P(x_0))) + KL(Q(e_1 || P(x_1 || P(x_0))) + ... +

KL(Q(e_t || P(x_t || P(x_{t-1})))) || P(x_t || P(x_{t-1}))).
\]
Since $e_0 = x_0 - s_0$ using (12), from the monotonicity and chain rule properties of KL divergence we have

$$KL(Q|P(x_0)) \leq KL(Q(e_0, x_1, y_1)|P(x_0, x_1, y_1)) = KL(Q(x_1, y_1)|P(x_1, y_1)) + KL(Q(e_0|x_1, y_1)|P(x_0|x_1, y_1)).$$

(27)

Given $x_1$ and $y_1$, the distribution of $x_0$ and $e_0$ are Gaussian with means $f(x_1) + B\pi(G(x_1), y_{s_1}^a)$ and $f(x_1) + B\pi(G(x_1), y_{s_1}^a) - s_0$ and covariance $\Sigma_w$, respectively. On the other hand, it holds that $KL(Q(x_1, y_1)|P(x_1, y_1)) = 0$. Therefore, we have

$$KL(Q(e_0)|P(x_0)) \leq s_0^T\Sigma_w^{-1}s_0.$$

(28)

For the following terms, we need to find $KL(Q(\gamma_{t}^a|e_t)|P(y_{t}^a|x_t))$ and $KL(Q(e_t|e_{t-1}, \gamma_{t-1}^a)|P(x_t|x_{t-1}, y_{t-1}^a))$. We have

$$KL(Q(\gamma_{t}^a|e_t)|P(y_{t}^a|x_t)) = (C_s(e_t - x_t))^T\Sigma_v^{-1}(C_s(e_t - x_t)) = (C_s r_t)^T\Sigma_v^{-1}(C_s r_t),$$

(29)

$$KL(Q(e_t|e_{t-1}, \gamma_{t-1}^a)|P(x_t|x_{t-1}, y_{t-1}^a)) = r_t^T\Sigma_w^{-1}r_t.$$  

(30)

Now, it holds that

$$KL(Q(e_0, \gamma_{t}^a, e_1, y_{t}^a, ..., e_t, y_{t}^a)|P(x_0, y_{0}^a, x_1, y_{1}^a, ..., x_t, y_{t}^a)) \leq$$

$$\leq s_0^T\Sigma_w^{-1}s_0 + \sum_{i=1}^{t}r_i^T\Sigma_w^{-1}r_i + \sum_{i=1}^{t}r_i^T(C_s^T\Sigma_v^{-1}C_s)r_i$$

$$\leq s_0^T\Sigma_w^{-1}s_0 + \sum_{i=1}^{t}\lambda_{\text{max}}(C_s^T\Sigma_v^{-1}C_s + \Sigma_w^{-1})\|r_i\|^2,$$

where we used the norm property $x^TQx \leq \lambda_{\text{max}}(Q)\|x\|^2$. Now, if $T(\alpha + b_x + b, s_0) < t_1$, then $\sum_{i=1}^{T(\alpha + b_x + b, s_0)}\|r_i\|^2 \leq \min\{T(\alpha + b_x + b, s_0), \sqrt{\frac{c_2}{c_1}}\frac{e^{-\beta}}{1-e^{-\beta}}\}\|s_0\|^2$ with probability $\delta(T(\alpha + b_x + b, s_0), b_x)$. However, if $T(\alpha + b_x + b, s_0) \geq t_1$ then $\sum_{i=1}^{T(\alpha + b_x + b, s_0)}\|r_i\|^2 \leq \min\{t_1, \sqrt{\frac{c_2}{c_1}}\frac{e^{-\beta}}{1-e^{-\beta}}\}\|s_0\|^2 + (T(\alpha + b_x + b, s_0) - t_1)b$ with probability $\delta(T(\alpha + b_x + b, s_0), b_x)$. Using the inequalities (24), (25) and (30), it holds that

$$KL(Q(y_{t}^a : y_{T(\alpha + b_x + b, s_0)}^a)|P(y_0 : y_{T(\alpha + b_x + b, s_0)})) \leq \lambda_{\text{max}}(\Sigma_w^{-1})\|s_0\|^2 +$$

$$\lambda_{\text{max}}(C_s^T\Sigma_v^{-1}C_s + \Sigma_w^{-1}) \max \left\{ \min \left\{ T(\alpha + b_x + b, s_0), \sqrt{\frac{c_2}{c_1}}\frac{e^{-\beta}}{1-e^{-\beta}} \right\}\|s_0\|^2, \right.$$

$$\left. \min \left\{ t_1, \sqrt{\frac{c_2}{c_1}}\frac{e^{-\beta}}{1-e^{-\beta}} \right\}\|s_0\|^2 + (T(\alpha + b_x + b, s_0) - t_1)b \right\} = b_c,$$

which means the system is $(\epsilon, \alpha)$-successful attackable with probability of $\delta(T(\alpha + b_x + b, s_0), b_x)$ with $\epsilon = \sqrt{1 - e^{-\beta}}$. 

From (5), $c_3$ can be viewed as a measure of the closed-loop system stability (larger $c_3$ means the system is ‘more’ stable); on the other hand, from Theorem 11, closed-loop perception-based systems with larger $c_3$ are more vulnerable to stealthy attacks as the conditions of the theorem are easier to satisfy. However, if the plant’s dynamics is very unstable, $T(\alpha + b_x + b, s_0)$ is smaller for a fixed $\alpha$ and $s_0$. Thus, the probability of attack success $\delta(T(\alpha + b_x + b, s_0), b_x)$ is larger for a fixed $b_x$. Moreover, in the extreme case when $b_c = 0$ (i.e., the attacker can exactly estimate the
plant state), the condition \( L_2b_\xi < \frac{c_2-(L_1+L_3\|B\|)c_4}{c_4}\sqrt{\frac{c_1}{c_2}} \theta r \) will be relaxed and the other condition \( L_1 + L_3\|B\| < \frac{c_3}{c_4} \) becomes less restrictive as \( L_1 \) becomes smaller. Therefore, in this case, if the attacker initiates the attack with arbitrarily small \( s_0 \), then \( \epsilon \) can be arbitrarily close to zero and the attack will be very close to being strictly stealthy. Hence, we the following result holds.

**Corollary 13** Assume \( b_\xi = 0, L_1 + L_3\|B\| < \frac{c_3}{c_4} \) with \( L_1 = L_1'(b_x + d) \) and \( L_3 = L_3'(b_x + d) \). Then, if \( f \in U_\rho \) with \( \rho = 2L_1'(b_x + b) \), the system (1) is \((\epsilon, \alpha)\)-attackable with probability \( \delta(T(\alpha + b_x + b, s_0), b_x) \), where \( \epsilon = \sqrt{1 - e^{-b_r}} \) for \( b_r = \left( \lambda_{\max}(\Sigma_w^{-1}) + \lambda_{\max}(G_s^T\Sigma_w^{-1}C_s + \Sigma_w^{-1}) \right) \times \min\{T(\alpha + b_x + b, s_0), \sqrt{\frac{c_2}{c_1} \frac{e-\beta}{1-e-\pi}} \} \|s_0\|^2 \).

Finally, the above results depend on the determining \( \rho \) such that \( f \in U_\rho \). Hence, the following result provides a sufficient condition for \( f \in U_\rho \).

**Proposition 14** Let \( V : \mathbb{R}^n \to \mathbb{R} \) be a continuously differentiable function satisfying \( V(0) = 0 \) and define \( U_{r_1} = \{ x \in B_{r_1} \mid V(x) > 0 \} \). Assume that \( \|\frac{\partial V(x)}{\partial x}\| \leq \beta(\|x\|) \) and for any \( x \in U_{r_1} \), it holds \( V(f(x)) - V(x) \geq \alpha(\|x\|) \), where \( \alpha(\|x\|) \) and \( \beta(\|x\|) \) are in class of functions Khalil (2002). Further assume that \( r_1 \) can be chosen arbitrarily large. Now, if \( \lim_{\|x\| \to \infty} \frac{\alpha(\|x\|)}{\beta(\|x\|)} = \gamma \), then \( f \in U_\rho \) for any \( \rho < \gamma \).

**Proof** Here, we prove the first case, and the second case proof follows similar steps. Since \( \lim_{\|x\| \to \infty} \frac{\alpha(\|x\|)}{\beta(\|x\|)} \to \infty \), there exists a bounded ball with radius \( r_2 \) (we denote it \( B_{r_2} \)) such that for all \( x \in S \) with \( S = \{ U_{r_1} - B_{r_2} \} \), it holds that \( \lambda(\|x\|)^2 \geq \rho \). Since the function \( V \) is differentiable, using the Mean-value theorem for the dynamics \( x_{t+1} = f(x_t) + d_t \) with \( \|d_t\| \leq \rho \) and for any \( x_t \in U_{r_1} \), we have

\[
V(x_{t+1}) - V(x_t) = V(f(x_t) + d_t) - V(x_t) = V(f(x_t)) + \frac{\partial V(x)}{\partial x} d_t - V(x_t) \geq \alpha(\|x\|) - \beta(\|x\|) \rho.
\]

(32)

Therefore, for any \( x_t \in S \) we have \( V(x_{t+1}) - V(x_t) > 0 \). Let define \( \eta = \min\{V(x_{t+1}) - V(x_t) \mid x_t \in S \) and \( V(x_t) \geq \alpha_{r_2} \}, \) where \( \alpha_{r_2} = \max_{x \in \partial B_{r_2}} V(x) \) for any \( r > 0 \). Such minimum exists as the considered set is compact and we have \( \eta > 0 \) in \( U_{r_1} \). Let us also assume \( x_0 = \{ x \in B_{r_2} \mid V(x) = \alpha_{r_2} \} \). Now, we claim that the trajectories starting from \( x_0 \) should leave the set \( U_{r_1} \) through the boundaries of \( B_{r_1} \). To show this, we know that for any \( x_t \in S \), \( V(x_t) \geq \alpha_{r_2} \) since \( V(x_{t+1}) - V(x_t) \geq \eta > 0 \). Then, for any \( t > 0 \)

\[
V(x_t) \geq V(x_0) + t \eta = \alpha_{r_2} + \eta t.
\]

(33)

The above inequality shows that \( x_t \) cannot stay in the set \( S \) forever as \( V(x) \) is bounded on the compact set \( S \). On the other hand, \( x_t \) cannot leave the set \( S \) through the boundaries satisfying \( V(x) = 0 \) or the surface of \( B_{r_2} \) because \( V(x_t) > \alpha_{r_2} \). Therefore, the trajectories should leave the set \( S \) through the surface of \( B_{r_1} \) and as \( r_1 \) can be chosen arbitrarily large, the trajectories of \( x_t \) will diverge to infinity.

Note that our results only focus on the existence of perception measurements \( G(x_t - s_t) \), obtained by shifting the current perception scene by \( s_t \) that results in \((\epsilon, \alpha)\)-successful attack, and not how to compute it. Further, to derive attack sequence using Attack Strategy I, the attacker needs
the estimation of the plant states. Thus, Attack Strategy II relaxes this assumption, with the attacker only needing to have knowledge about the plant’s (open-loop) dynamics \( f \) and the computation power to calculate \( s_{t+1} = f(s_t) \) ahead of time.

### 3.2. Attack Strategy II: Using Plant Dynamics

Similarly to Attack Strategy I, consider the attack sequence where \( z_t^a \) and \( y_t^{s,a} \), for all \( t \geq 0 \), satisfy

\[
z_t^a = G(x_t^a - s_t), \quad y_t^{s,a} = C_s(x_t^a - s_t) + v_t^{s,a}, \quad s_{t+1} = f(s_t),
\]

for some nonzero \( s_0 \). However, here the attacker does not need plant state estimate; they simply follow plant dynamics \( s_{t+1} = f(s_t) \) to find the desired measurements’ perturbations. Now, we define the state \( e_t \triangleq x_t^a - s_t \) as the fake state, and the attacker’s goal is to make the system believe that the plant state is equal to \( e_t \).

**Theorem 15** Consider the system (1) with closed-loop control as in Assumption 2. Assume that both functions \( f' \) and \( \pi' \) are locally Lipschitz in the set \( \mathcal{D} \), with constants \( L_f' \) and \( L_{\pi}' \), respectively. Moreover, assume \( b_x \) has the maximum value such that the inequalities \( L_1 + L_3\|B\| < \frac{c_3}{c_4} \) and \( L_2 b_x < \frac{c_3-(L_1+L_3\|B\|)c_3}{c_4} \frac{c_3}{c_2} \theta \) with \( 0 < \theta < 1 \) are satisfied, where \( L_2 = L_f'(\alpha + b_x) \), \( L_1 = L_f'(\alpha + d) \) and \( L_3 = L_{\pi}'(b_x + d) \). Then, the system (1) is \((\epsilon, \alpha)\)-attackable with probability \( \delta(T(\alpha + b_x + b, s_0), b_x) \) and \( b = \frac{c_3}{c_3-(L_1+L_3\|B\|)c_3} \frac{c_3}{c_1} \|B\| \theta \), for some \( \epsilon > 0 \), if \( f \in U_0 \).

**Proof** We need to show that the sequence of perturbed image and sensor values \( \{z_0^a, y_0^{s,a}\}, \{z_1^a, y_1^{s,a}\}, \ldots \) obtained by Attack Strategy II are \((\epsilon, \alpha)\)-successful attack. Again, by defining \( e_t = x_t^a - s_t \ flt} \) and \( r_t = e_t - x_t \), we get \( z_t^a = G(e_t), y_t^{s,a} = C_s e_t + v_t^{s,a} \) and

\[
e_{t+1} = f(x_t^a) - f(s_t) + B\pi(G(e_t), y_t^{s,a}) + w_t^a,
\]

\[
r_{t+1} = f(r_t) + B\pi(G(r_t), C_s r_t) + f(x_t^a) - f(s_t) - f(x_t) - f(r_t),
\]

\[
+ B\pi(G(e_t), y_t^{s,a}) - B\pi(G(x_t), y_t^{s,a}) - B\pi(G(r_t), C_s r_t) = h(r_t) + \sigma_1 + \sigma_2.
\]

with \( \sigma_1 = f(x_t^a) - f(s_t) - f(x_t) - f(r_t) \) and \( \sigma_2 = B\pi(G(e_t), y_t^{s,a}) - B\pi(G(x_t), y_t^{s,a}) - B\pi(G(r_t), C_s r_t) \).

Using the Mean value theorem and equality \( x_t^a = r_t + s_t + x_t \) from equation (12) we have

\[
f(x_t^a) = f(x_t + s_t + r_t) = f(s_t) + \frac{\partial f}{\partial x}|_{\Theta(s,s+x+r)}(x_t + r_t),
\]

\[
f(r_t) = f(0) + \frac{\partial f}{\partial x}|_{\Theta(0,r)}r_t = \frac{\partial f}{\partial x}|_{\Theta(0,r)}r_t,
\]

\[
f(x_t) = f(0) + \frac{\partial f}{\partial x}|_{\Theta(0,s)}x_t = \frac{\partial f}{\partial x}|_{\Theta(0,s)}x_t.
\]

Therefore, it holds that

\[
\sigma_1 = (\frac{\partial f}{\partial x}|_{\Theta(s,s+x+r)} - \frac{\partial f}{\partial x}|_{\Theta(0,r)})x_t + (\frac{\partial f}{\partial x}|_{\Theta(s,s+x+r)} - \frac{\partial f}{\partial x}|_{\Theta(0,r)})r_t.
\]

Similarly, using the Mean Value theorem and \( e_t = r_t + x_t \) we have

\[
\pi(G(e_t), y_t^{s,a}) = \pi(G(x_t), y_t^s) + \frac{\partial \pi}{\partial x}|_{\Theta(x,x+r)}r_t,
\]

\[
\pi(G(r_t), C_s r_t) = \frac{\partial \pi}{\partial x}|_{\Theta(0,r)}r_t.
\]
Hence, we get

\[ \sigma_2 = B_1 \frac{\partial \pi}{\partial x} \Theta_{(x,x+\epsilon)} - \frac{\partial \pi}{\partial x} \Theta_{(0,r)} r_t \]  

(39)

Since the function \( \frac{\partial f}{\partial x}(x) \) is Lipschitz, for all \( 0 \leq t \leq T(\alpha + b_x + b, s_0) \) we have \( \| \frac{\partial f}{\partial x} \Theta_{(x,x+\epsilon)} - \frac{\partial f}{\partial x} \Theta_{(0,x+\epsilon)} \| \leq L_f' \| x^2 \| + \| x_t \| \) \) and \( \| \frac{\partial f}{\partial x} \Theta_{(x,x+\epsilon)} - \frac{\partial f}{\partial x} \Theta_{(0,x+\epsilon)} \| \leq L_f' \| x^2 \| + \| r_t \| \) \). Therefore, \( \| \sigma_1 \| \leq L_2 b_x + L_1 \| r_t \| \). Similarly, \( \| \sigma_1 \| \leq L_2 b_x + L_1 \| r_t \| \) and following the same procedure as Theorem 11 we have

\[ \mathbb{P}(\| \sigma_1 \| \leq L_2 b_x + L_1 \| r_t \|) \leq \mathbb{P}(\| s_0 \| \leq \frac{c_0}{c_4} e^{-c_1} \| r_0 \|) \]

using Lemma 12 for all \( \| r_0 \| = \| s_0 \| < \sqrt{c_2} \), there exists \( t_1 > 0 \), such that for all \( t \geq t_1 \) we have \( \| r_t \| \leq b \) with \( b = \frac{c_4}{c_3-L_3|B|c_4} \sqrt{c_2} \frac{L_2 b_x}{c_1} \) and for all \( 0 \leq t < t_1 \) we have \( \| r_t \| \leq \sqrt{c_2} e^{-c_1} \| s_0 \| \)

with \( \beta = \frac{(1-\epsilon)}{c_3-L_3|B|c_4} \). On the other hand, the (deterministic) dynamics \( s_{t+1} = f(s_t) \) with nonzero \( s_0 \) will reach to \( \| s_t \| \geq \alpha + b_x + b \) for some \( t \geq T(\alpha + b_x + b, s_0) \) as \( f \in U_0 \). Using reverse triangle inequality we can get

\[ \| s_t \| - \| x^2 \| \leq \| x^{a} - s_t \| = \| e_t \| = \| x_t + r_t \| \]

\[ \leq b_x + b \Rightarrow -b - b_x + b + b_x + \alpha = \alpha \leq \| x^2 \| \]

(40)

with probability \( \delta(T(\alpha + b_x + b, s_0), b_x) \). Now, we need to show that the designed \( y_t^a \) satisfies the stealthiness condition. In other words, we need to show that \( KL(Q(y_0^a : y_{T(T(\alpha + b_x + b, s_0))})||P(y_0 : y_{T(\alpha + b_x + b, s_0)})) \leq \log(\frac{1}{1-\epsilon}) \) for some \( \epsilon > 0 \). Since the sequences \( y_0^a, \ldots, y_t^a \) and \( y_0, \ldots, y_t \) are generated by \( e_0, \ldots, e_t \) \( x_0, \ldots, x_t \), respectively, using the Data-processing inequality of divergence and following the same procedure as Theorem 11 we have

\[ KL(Q(y_0^a : y_{T(\alpha + b_x + b, s_0)}))||P(y_0 : y_{T(\alpha + b_x + b, s_0)})) \leq s_0^T \Sigma_w^{-1} s_0 + \sum_{i=1}^{t} \lambda_{\max}(C_s^T \Sigma_u^{-1} C_s + \Sigma_u^{-1}) \| r_i \|^2. \]

(41)

Similar argument as in Theorem 11 results in

\[ KL(Q(y_0^a : y_{T(\alpha + b_x + b, s_0})))||P(y_0 : y_{T(\alpha + b_x + b, s_0})) \leq \lambda_{\max}(\Sigma_w^{-1}) \| s_0 \|^2 + \lambda_{\max}(C_s^T \Sigma_u^{-1} C_s + \Sigma_u^{-1}) \max \left\{ \min \left\{ T(\alpha + b_x + b, s_0), \frac{\sqrt{c_2}}{c_1} e^{-\beta} \right\} \| s_0 \|^2, \right\} \]

\[ \min \left\{ t_1, \frac{\sqrt{c_2}}{c_1} e^{-\beta} \right\} \| s_0 \|^2 + (T(\alpha + b_x + b, s_0) - t_1) b \}

(42)

which means the system is \((\epsilon, \alpha, \alpha)\)-successful attackable with probability of \( \delta(T(\alpha + b_x + b, s_0), b_x) \) with \( \epsilon = \sqrt{1 - e^{-c_1}} \). \( \blacksquare \)

Unlike in Theorem 11, \( L_1 \) and \( L_2 \) in Theorem 15 increase as \( \alpha \) increases. Therefore, unless \( L_f' = 0 \), one cannot claim that the attack can be \( \epsilon \)-stealthy for arbitrarily large \( \alpha \) as the inequality \( L_1 + L_3|B| \leq \frac{c_0}{c_4} \) might not be satisfied. However, in an extreme case where the system is linear, it holds that \( L_f' = 0 \) and \( L_1 = L_2 = 0 \). Also, for linear-time invariant (LTI) systems, Attack Strategies I and II become identical as \( s_{t+1} = A(\hat{x}^a_t) - A(\hat{x}^a_t - s_t) = A s_t \). Hence, the following holds.
Corollary 16 Consider an LTI perception-based control system with \( f(x_t) = Ax_t \). If \( L_{3\max}\|B\| < \frac{c_2}{c_1} \) with \( L_3 = L_{2\max}(b_x + d) \), and the matrix \( A \) is unstable, then the system is \((\epsilon, \alpha)\)-attackable with probability \( \delta(T(\alpha, s_0), b_x) \), for arbitrarily large \( \alpha \) and \( \epsilon = \sqrt{1 - e^{-b_x}} \), where \( b_x = \left( \lambda_{max}(\Sigma^{-1}_w) + \lambda_{max}(C^T\Sigma^{-1}_vC_s + \Sigma^{-1}_w) \times \sqrt{\frac{c_2}{c_1}} \frac{e^{-\beta}}{1 - e^{-\beta}} \right)\|s_0\|^2 \) and \( e^{-\beta} \) is the largest eigenvalue of the closed-loop control system.

Both Theorem 11 and Theorem 15 assume end-to-end controller that directly maps the perception and sensor measurements to the control input. However, there exist controllers that first extract the state information using the perception module \( P \) and then use a feedback controller to find the control input Dean et al. (2020a,b); Dean and Recht (2020). For instance, for a linear feedback controller denoted by \( u_t = \pi(y_t) = K y_t \), we have \( L_{\pi\alpha} = 0 \). Thus, we can obtain the following.

Corollary 17 Consider the LTI perception-based control system with \( f(x_t) = Ax_t \) and a linear feedback controller. If the matrix \( A \) is unstable, the system is \((\epsilon, \alpha)\)-attackable with probability one for arbitrarily large \( \alpha \) and \( \epsilon = \sqrt{1 - e^{-b_x}} \), where \( b_x = \left( \lambda_{max}(\Sigma^{-1}_w) + \lambda_{max}(C^T\Sigma^{-1}_vC_s + \Sigma^{-1}_w) \times \sqrt{\frac{c_2}{c_1}} \frac{e^{-\beta}}{1 - e^{-\beta}} \right)\|s_0\|^2 \) and \( e^{-\beta} \) is the largest eigenvalue of the closed-loop control system.

Finally, note that for an LTI system, the attacker will be impactful if and only if \( s_0 \) is not orthogonal to all unstable eigenvectors of the matrix \( A \). Therefore, by choosing such \( s_0 \) that is also arbitrarily close to zero the attacker can be \( \epsilon \)-stealthy with \( \epsilon \) to be near zero.

4. Simulation Results

We illustrate and evaluate our vulnerability analysis of perception-based controller systems on a case-study. Specifically, we consider a fixed-base inverted pendulum equipped with an end-to-end controller and a perception module that estimates the pendulum angle from camera images. By using \( x_1 = \theta \) and \( x_2 = \dot{\theta} \), the inverted pendulum dynamics can be modeled in the state-space form

\[
\dot{x}_1 = x_2, \\
\dot{x}_2 = g \frac{\sin x_1}{r} - \frac{b}{mr^2} x_2 + \frac{L}{mr^2}.
\]

Here, \( \theta \) is the angle of pendulum rod from the vertical axis measured clockwise, \( b \) is the Viscous friction coefficient, \( r \) 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 Formal’ski (2006). Finally, we assumed \( g = 9.8 \), \( m = .2Kg \), \( b = .1 \), \( r = .3m \) and discretized the model with \( T_s = 10 \ m/s \).

Using Lyapunov’s indirect method one can show that the origin of the above system is unstable because the linearized model has an unstable eigenvalue. However, the direct Lyapunov method can help us to find the whole unstable region \(-\pi < \theta < \pi \) (see Khalil (2002)). We used a data set \( S \) with 500 sample of pictures of the fixed-base inverted pendulum with different angles in \((-\pi, \pi)\) to train a DNN \( P \) (perception module) to estimate the angle. Fig. 2 shows the predicted values of the

![Figure 2: Perception map (P) predicted value vs true values of pendulum angle \( \theta \).](image-url)
We first choose $s_0 = [0.01 \ 0.01]^T$ and used Attack Strategy I to design attacks. To derive the current adversarial image at each time step the attacker receives the actual image and compromise it by deviating the pendulum rod by $s_t$ degrees. This compromised image is used by the perception module to evaluate the system state. Again, the attacker does not need to have access to the perception map $P$; the knowledge about the dynamics $f$ and estimate of the current plant state $\hat{x}_t^a$ is sufficient to craft the perturbed images. Fig. 3(a) shows the actual pendulum pod angle for different estimation uncertainty level $b_\zeta$ (by the attacker) when the attack starts at $t = 0$. In both cases, the attacker can drive the pendulum pod into an unsafe region. Fig. 3(b) shows the residue signal over time; the attack stealthiness level decreases as $b_\zeta$ increases, consistent with our results in Sec. 3.

In Fig. 4(b), the residue of the system in normal condition as well as under Attack Strategy II. We can see that the residue level of both Attack Strategy I with $b_\zeta = 0.05$ and Attack Strategy II are the same as for the system in normal condition. The red and blue line in Fig. 4(a) also show the pendulum pod angle trajectory for Attack Strategy II and normal condition, respectively.

5. Conclusion

In this work, we considered the problem of resiliency under sensing and perception attacks for perception-based control systems, focusing on a general class of nonlinear dynamical plants. We assumed that the noiseless closed-loop system equipped with an end-to-end controller and anomaly.
detector, is exponentially stable on a set around the equilibrium point. We introduced the notion of $\epsilon$-stealthiness as a measure of difficulty in attack detection from the set of perception measurements and sensor values. Further, we derived sufficient conditions for an effective yet $\epsilon$-stealthy attack sequence to exist. Here, control performance degradation was considered as moving the system state outside of the safe region defined by a bounded ball with radius $\alpha$, resulting in an $(\epsilon, \alpha)$-successful attack. Finally, we illustrated our results on fixed-base inverted pendulum case-study.

References


