2016 IEEE 55th Conference on Decision and Control (CDC)
ARIA Resort & Casino
December 12-14, 2016, Las Vegas, USA

When Adversary Encounters Uncertain Cyber-physical Systems: Robust Zero-dynamics Attack with Disclosure Resources

Gyunghoon Park, Hyungbo Shim, Chanhwa Lee, Yongsoon Eun, and Karl H. Johansson

***Abstract*— In this paper we address the problem of designing a robust stealthy attack for adversaries to compromise an uncertain cyber-physical system without being detected. We first re-interpret the zero-dynamics attack based on the normal form representation. Then, a new alternative zero dynamics attack is presented for *uncertain* systems. This alternative employs a disturbance observer and does not require exact system knowledge in order to remain stealthy. The proposed robust zero-dynamics attack needs a nominal model of the system and, in addition, utilizes the input and output signals of the system. The proposed attack illustrates how the adversary is able to use disclosure resources instead of exact model knowledge. A simulation result with a hydro-turbine power system is presented to verify the attack performance.**

I. Introduction

Nowadays modern control systems integrate computing devices, physical plants, and communication networks. Such a cyber-physical system (CPS) is a promising framework for cost efficiency and high productivity. The cyber-physical approach has gone beyond the fundamental limitations of the conventional methodologies and has achieved a great success in numerous industrial fields [1], [2].

At the same time, CPSs are more threatened by attacks, as their network connections are easier to be accessed for anonymous users. Serious outbreaks of malicious cyber threats already have been reported in recent years. Some remarkable instances include the attacks on the U.S. electric grid [3] and the Stuxnet malware [4]. As a natural consequence, the security of CPS has attracted widespread attention with emerging resilient control and secure estimation schemes [5]–[8].

With such increased interest on the cyber-security, a variety of attack scenarios, such as denial-of-service (DoS) attack, replay attack [9], zero-dynamics attack [5], [6], [10], bias injection attack [5], and so on, have been studied from a control-theoretic perspective. As highlighted in most of these works, *stealthiness* is of utter importance for success of the adversary; that is, when an attack signal enters a CPS, its impact should not be captured by sensor and anomaly detector. The zero-dynamics attack is a systematic strategy to be undetected and simultaneously to inject a large amount of false data into the plant. Its stealthy property mainly comes from that the adversary duplicates the real unstable zero dynamics of non-minimum phase physical plants, so the attack signal conceals itself in the so-called output-nulling space [5], [10]. However, the attack design heavily relies on model knowledge. Lack of system information thus leaves the attack revealed, which means that it becomes not stealthy anymore for such uncertain plants [10]. If so, can we be safe from those stealthy attacks simply because model uncertainty commonly exists for most physical systems?

Interestingly, we find in this paper that it may not be the case when the attacker employs robust control techniques in their attacks. Specifically, we solve the problem of constructing *robust* zero-dynamics attack that is stealthy for uncertain non-minimum phase plants. Moving away from the traditional methods, the underlying idea is to construct an auxiliary zero dynamics which will replace the role of real zero-dynamics, so that the real zero-dynamics is left alone. Then, the state components corresponding to its unstable mode will diverge, which is not observed by the output. This idea is implemented by representing the influence of model uncertainty and the real zero-dynamics into a (so-called) lumped disturbance, and by constructing the robust controller which estimates and compensates the lumped disturbance. All this is actually done by the disturbance observer [12], [13]. The price to pay for less model knowledge is the necessity of the control input and the plant's output information,

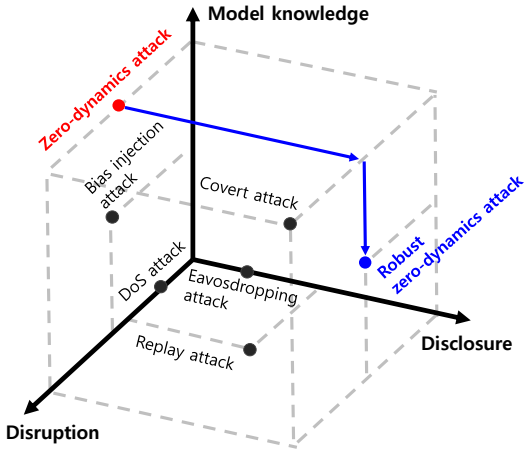


Fig. 1. Cyber-physical attack space [8] with model knowledge, disruption, and disclosure resources: The robust zero-dynamics attack is at entirely new location.

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (B0101-16-0557, Resilient Cyber-Physical Systems Research).

G. Park, H. Shim, and C. Lee are with ASRI, Department of Electrical and Computer Engineering, Seoul National University, Korea. Y. Eun is with Department of Information & Communication Engineering, Daegu Gyeongbuk Institute of Science & Technology, Korea. K. H. Johansson is with the ACCESS Linnaeus Centre, School of Electrical Engineering, KTH Royal Institute of Technology, Sweden.

and so the proposed robust zero-dynamics attack requires more *disclosure resources* [8], as depicted in Fig. 1.

Notation: For column vectors a and b, we write $[a;b]$ for $\mathrm{col}(a,b) = [a^\intercal, b^\intercal]^\intercal$.

II. Re-interpretation of Zero-dynamics Attack with Normal Form

Zero-dynamics attack is a systematic methodology for attacking a class of cyber-physical systems (CPSs) having unstable zeros [5], [6], [10]. The basic concept is that, as its name implies, the attack generator maliciously disguises as the unstable zero-dynamics of the plant and injects its diverging output through the actuator channel. This consequently leads to the feature that the actual (zero-dynamics) state grows as time goes by, while being close to the output-nulling space (so the corresponding output is almost zero, by which we call it a *stealthy attack*).

Although geometric control framework has been employed as a tool for its analysis [5], we present in this section another way to interpret the zero-dynamics attack for the purpose of gaining further insight. The new interpretation is based on the *Byrnes-Isidori normal form*, in which a single-input single-output (SISO) physical system is represented as[1]

$$\dot{z} = Sz + GC_\nu x, \tag{1a}$$
$$\dot{x} = A_\nu x + B_\nu\big(\psi^\intercal z + \phi^\intercal x + g(u_\mathsf{c} + a)\big), \tag{1b}$$
$$y = C_\nu x \tag{1c}$$

where $u_\mathsf{c} \in \mathbb{R}$ is the control input, $y \in \mathbb{R}$ is the measurement output, $z \in \mathbb{R}^{n-\nu}$ and $x \in \mathbb{R}^\nu$ are the states, and $a \in \mathbb{R}$ is the attack signal. For an integer i, the matrices A_i, B_i, and C_i are defined by

$$A_i := \begin{bmatrix} 0_{i-1} & I_{i-1} \\ 0 & 0_{i-1}^\intercal \end{bmatrix}, \quad B_i := \begin{bmatrix} 0_{i-1} \\ 1 \end{bmatrix}, \quad C_i := \begin{bmatrix} 1 & 0_{i-1}^\intercal \end{bmatrix}$$

where $0_i \in \mathbb{R}^i$ is a zero vector, and S, G, ψ, ϕ, and g are constant matrices of appropriate size. The control input u_c in (1) is supposed to be generated by *a priori* given output feedback controller

$$\dot{c} = Ec + H(y_r - y), \quad u_\mathsf{c} = Jc + K(y_r - y) \tag{2}$$

where $c \in \mathbb{R}^{n_\mathsf{c}}$ is the controller state, and $y_r \in \mathbb{R}$ is the reference signal. The reference $y_r : \mathbb{R} \to \mathbb{R}$ is assumed to be sufficiently smooth and bounded, and their derivatives are also bounded.

In what follows, we pay our attention to a class of non-minimum phase systems with *hyperbolic*[2] zero-dynamics. Then, we may assume the following without loss of generality (by applying a suitable coordinate change for z).

[1]Any SISO linear system can be expressed as the Byrnes-Isidori normal form (1) [11, Chapter 13]. In this form, the zeros of the transfer function of the SISO linear system coincide with the eigenvalues of S. Hence, $\dot{z} = Sz$ is called the *zero-dynamics*.

[2]By *hyperbolic* zero-dynamics, we mean there is no zero on the imaginary axis of the complex plane. This is requested because the role of the attack to be presented is (as we shall explain precisely) to leave the zero-dynamics alone, and those marginal modes of the zero-dynamics are not enough to destabilize the system.

Assumption 1: The z-dynamics (1a) has the form of

$$\begin{bmatrix} \dot{z}_u \\ \dot{z}_s \end{bmatrix} = \begin{bmatrix} S_u & 0 \\ 0 & S_s \end{bmatrix} \begin{bmatrix} z_u \\ z_s \end{bmatrix} + \begin{bmatrix} G_u \\ G_s \end{bmatrix} C_\nu x \tag{3}$$

where $S_u \in \mathbb{R}^{n_u \times n_u}$ and $S_s \in \mathbb{R}^{n_s \times n_s}$ are anti-Hurwitz and Hurwitz, respectively, $n_u \geq 1$, and $z = [z_u; z_s] \in \mathbb{R}^{n_u+n_s}$. □

Under Assumption 1, the (non-robust) zero-dynamics attack is commonly constructed as

$$\dot{z}^\mathsf{a} = Sz^\mathsf{a}, \quad a_\mathrm{za} = -\frac{1}{g}\psi^\intercal z^\mathsf{a} \tag{4}$$

where $z^\mathsf{a} =: [z_u^\mathsf{a}; z_s^\mathsf{a}] \in \mathbb{R}^{n_u+n_s}$. (The superscript 'a' is used to indicate signals that are generated by the adversary.)

For the analysis, we introduce the attack-free closed-loop system

$$\begin{aligned} \dot{z}_0 &= Sz_0 + GC_\nu x_0, \\ \dot{x}_0 &= A_\nu x_0 + B_\nu\big(\psi^\intercal z_0 + \phi^\intercal x_0 + g u_{\mathsf{c},0}\big), \\ \dot{c}_0 &= Ec_0 + H\big(y_r - C_\nu x_0\big), \quad u_{\mathsf{c},0} = Jc_0 + K\big(y_r - C_\nu x_0\big) \end{aligned} \tag{5}$$

(which is derived by putting $a(t) \equiv 0$ into (1)). The nature of the zero-dynamics attack (4) is then reinterpreted in the following proposition. (Hereinafter, without loss of generality let $t = 0$ be the time when the attack $a(t)$ enters the system.)

Proposition 1: Suppose that Assumption 1 holds and the attack-free closed-loop system (5) is asymptotically stable. Then the closed-loop system (1)–(4) under the zero-dynamics attack $a = a_\mathrm{za}$ satisfies the following:

(a) If $z_u^\mathsf{a}(0) \neq 0$, then

$$\|z_u(t)\| \to \infty \quad \text{as} \quad t \to \infty. \tag{6}$$

(b) There are $k > 0$ and $\lambda > 0$ such that

$$\big\|[x(t); c(t)] - [x_0(t); c_0(t)]\big\| \leq ke^{-\lambda t}\|z^\mathsf{a}(0)\| \tag{7}$$

with $[z_0(0); x_0(0); c_0(0)] = [z(0); x(0); c(0)]$. □

Proposition 1 indicates that, with non-zero $z_u^\mathsf{a}(0)$ but small $\|z^\mathsf{a}(0)\|$, the plant's (partial) state diverges while the (attacked) real output $y(t) = C_\nu x(t)$ remains close to the attack-free output $y_0(t) = C_\nu x_0(t)$ (by (7)). It implies that stealthy attack is achieved.

Proof: Consider $\tilde{z}^\mathsf{a} := z - z^\mathsf{a}$ which transforms the closed-loop system (1)–(4) into

$$\begin{aligned} \dot{\tilde{z}}^\mathsf{a} &= S\tilde{z}^\mathsf{a} + GC_\nu x, \\ \dot{x} &= A_\nu x + B_\nu\big(\psi^\intercal \tilde{z}^\mathsf{a} + \phi^\intercal x + g u_\mathsf{c}\big), \\ \dot{c} &= Ec + H\big(y_r - C_\nu x\big), \quad u_\mathsf{c} = Jc + K\big(y_r - C_\nu x\big). \end{aligned} \tag{8}$$

Notice that (8) is nothing but the very attack-free closed-loop system (5) whose initial condition is slightly perturbed by $\tilde{z}^\mathsf{a}(0) = z(0) - z^\mathsf{a}(0)$. Since (8) is asymptotically stable by the assumption, we have

$$\big\|[\tilde{z}^\mathsf{a}(t); x(t); c(t)] - [z_0(t); x_0(t); c_0(t)]\big\| \leq ke^{-\lambda t}\|z^\mathsf{a}(0)\|$$

with positive constants k and λ, which implies the item (b). On the other hand, whenever $z_u^\mathsf{a}(0) \neq 0$, the solution $z^\mathsf{a}(t)$

of the unstable system (4) must diverge as time goes on. The actual state $z(t) = z^{\mathsf{a}}(t) + \tilde{z}^{\mathsf{a}}(t)$ also does, while $\tilde{z}^{\mathsf{a}}(t)$ being bounded. ■

Remark 1: From the analysis, it is clear that lower order dynamics $\dot{z}^{\mathsf{a}}_u = S_u z^{\mathsf{a}}_u$ and $a_{\mathrm{za}} = -(1/g)\psi_u^{\mathsf{T}} z^{\mathsf{a}}_u$ (where ψ_u is a suitable partition of ψ) is enough to construct the zero-dynamics attack. □

We emphasize that full model knowledge on the plant (1) is necessary for the zero-dynamics attack. In practice, it is not always possible for the attacker (as well as for the defender) to obtain the exact information on the plant. Even small model uncertainty can make the zero-dynamics attack detectable, so the attack is not *stealthy* anymore (as studied in [10]). This finding raises a question about how to construct *robustly* stealthy attack for uncertain plants, which is the main topic of the next section.

III. Robust Zero-dynamics Attack for Uncertain Systems

A. Problem Formulation

In what follows, it is assumed that the plant (1) has the following model uncertainty.

Assumption 2: All the parameters S, G, ψ, ϕ, and g are uncertain, but belong to known[3] finite parameter intervals. In particular, $0 < \underline{g} \le g \le \overline{g}$ with $\underline{g}$ and $\overline{g}$ known. □

As mentioned above, we are interested in the problem of constructing a *robust zero-dynamics attack* against uncertain physical plants. Without full model knowledge, the adversaries may have to build their attack strategy with the following (attack-free) *nominal* plant of (1):

$$\dot{z}_{\mathsf{n}} = S_{\mathsf{n}} z_{\mathsf{n}} + G_{\mathsf{n}} C_\nu x_{\mathsf{n}}, \tag{9a}$$

$$\dot{x}_{\mathsf{n}} = A_\nu x_{\mathsf{n}} + B_\nu\big(\psi_{\mathsf{n}}^{\mathsf{T}} z_{\mathsf{n}} + \phi_{\mathsf{n}}^{\mathsf{T}} x_{\mathsf{n}} + g_{\mathsf{n}} u_{\mathsf{n}}\big), \tag{9b}$$

$$y_{\mathsf{n}} = C_\nu x_{\mathsf{n}} \tag{9c}$$

where $z_{\mathsf{n}} \in \mathbb{R}^{n-\nu}$ and $x_{\mathsf{n}} \in \mathbb{R}^{\nu}$ are the states, and $u_{\mathsf{n}} \in \mathbb{R}$ is the control input generated by the existing controller (2)

$$\dot{c}_{\mathsf{n}} = Ec_{\mathsf{n}} + H(y_r - y_{\mathsf{n}}), \quad u_{\mathsf{n}} = Jc_{\mathsf{n}} + K(y_r - y_{\mathsf{n}}). \tag{9d}$$

The scalar g_{n} and the matrices S_{n}, G_{n}, ψ_{n}, and ϕ_{n} stand for nominal components of g, S, G, ψ, and ϕ, respectively. We suppose that the nominal closed-loop system (9) is asymptotically stable.

Now, motivated by Proposition 1, we formulate the problem as follows.

Problem Statement: For given $\underline{z}_u > 0$ and $\epsilon > 0$, construct an attack generator

$$\dot{q} = \mathcal{F}(q, u_{\mathsf{c}}, y, t), \quad a = \mathcal{G}(q, u_{\mathsf{c}}, y, t) \tag{10}$$

that achieves the following properties simultaneously:

(a) $\|z_u(t)\|$ becomes larger than $\underline{z}_u > 0$ within a finite time $t = T$,

(b) the difference $\|y(t) - y_{\mathsf{n}}(t)\|$ is smaller than a threshold $\epsilon > 0$ until the attack succeeds (that is, $\forall t \in [0, T]$). □

[3]They are 'known to attackers.' The interval can be conservative so that the attacker can overestimate those intervals.

The item (a) indicates the ability of the attack to compromise the plant's internal state (where $\underline{z}_u$ is the attacker's choice), while the item (b) means stealthiness of the attack. It is noted that, compared with the traditional structure (4), the attack generator (10) makes use of the signals u_{c} and y. This is in fact the price to pay for the *robustness* against model uncertainty; that is, instead of using less model knowledge, the attacker relies more on the input and output information of the plant to adjust to uncertain environment on-line.

Assumption 3: The measurement output $y(t)$ and the control input $u_{\mathsf{c}}(t)$ are available to attackers. □

Remark 2: It is noticed that the item (b) in Problem Statement measures the difference between the output $y(t)$ under attack and, not the attack-free output $y_0(t)$ of the (actual) uncertain system (1) as in Proposition 1, but $y_{\mathsf{n}}(t)$ of its nominal counterpart (9). At a first glance, recalling that the *perfect* stealthiness is achieved when $y(t) \equiv y_0(t)$ (not $y(t) \equiv y_{\mathsf{n}}(t)$), the attack (10) with this alternative definition seems easily revealed if the model uncertainty is too large to neglect the difference of the actual and nominal plants' dynamics. Even for such large uncertainty, however, this is often not the case as long as the existing controller (2) is also robust against parametric uncertainties. Indeed, when a tracking or regulating problem is (robustly) solved by (2) for both actual and nominal systems (with no attack), their outputs $y_{\mathsf{n}}(t)$ and $y_0(t)$ reach the same reference $y_r(t)$ in the end. It means that $y_{\mathsf{n}}(t) \approx y_0(t)$ during the steady-state operation of the system, within which the attack is usually initiated. In summary, we claim in this paper that the proposed attack tends to be stealthy if the uncertainty is not large or if the attack enters the system in the steady state, whereas for the same situations the conventional zero-dynamics attack (4) is not stealthy (as aforementioned). We will come to this point in the simulation of Section IV again. ■

For convenience, let us denote the nominal state of (9) as $\chi_{\mathsf{n}} := [z_{\mathsf{n}}; x_{\mathsf{n}}; c_{\mathsf{n}}]$. Then (9) is rewritten by

$$\dot{\chi}_{\mathsf{n}} = \mathsf{A}_{\mathsf{n}}\chi_{\mathsf{n}} + \mathsf{B}_{\mathsf{n}} y_r, \quad y_{\mathsf{n}} = \mathsf{C}_{\mathsf{n}}\chi_{\mathsf{n}} \tag{11}$$

where

$$\mathsf{A}_{\mathsf{n}} := \begin{bmatrix} S_{\mathsf{n}} & G_{\mathsf{n}}C_\nu & 0 \\ B_\nu\psi_{\mathsf{n}}^{\mathsf{T}} & A_\nu + B_\nu\phi_{\mathsf{n}}^{\mathsf{T}} - g_{\mathsf{n}}B_\nu K C_\nu & g_{\mathsf{n}}B_\nu J \\ 0 & -HC_\nu & E \end{bmatrix},$$

$$\mathsf{B}_{\mathsf{n}} := \begin{bmatrix} 0_{n-\nu}; g_{\mathsf{n}}B_\nu K; H \end{bmatrix}, \quad \mathsf{C}_{\mathsf{n}} := \begin{bmatrix} 0_{n-\nu}^{\mathsf{T}} & C_\nu & 0_{n_{\mathsf{c}}}^{\mathsf{T}} \end{bmatrix}.$$

This system will play the role of a reference system for the analysis of attack performance, and so, its initial condition $[z_{\mathsf{n}}(0); x_{\mathsf{n}}(0); c_{\mathsf{n}}(0)] (= \chi_{\mathsf{n}}(0))$ will be regarded as the same as $[z(0); x(0); c(0)]$. Finally, we assume that the initial condition $[z(0); x(0); c(0)]$ of the system is bounded:

Assumption 4: There is a compact set $\mathcal{X}_0$ such that $[z(0); x(0); c(0)] \in \mathcal{X}_0$. □

We note that this assumption is not unrealistic because the size of $\mathcal{X}_0$ can be arbitrarily large.

B. Idea for Robust Zero-dynamics Attack

We introduce an attack policy to *robustly* compromise the internal z-dynamics. We note in advance that the method

to be presented here is not realizable yet, but we will shortly make it feasible in the next subsection. A rough attack scenario is that an artificial nominal zero-dynamics is constructed to substitute for the real zero-dynamics, while the real one becomes stand-alone detached from the closed-loop system. Then the state of the unstable (real) zero-dynamics without any control may diverge.

To see this, let us duplicate the nominal zero-dynamics (9a) using the output y of (1) as

$$\dot{z}_{\mathsf{n}}^{\mathsf{a}} = S_{\mathsf{n}} z_{\mathsf{n}}^{\mathsf{a}} + G_{\mathsf{n}} y \tag{12}$$

$$z_{\mathsf{n}}^{\mathsf{a}}(0) = z(0) \in \mathcal{Z}_0 \tag{13}$$

where $\mathcal{Z}_0$ is the projection of $\mathcal{X}_0$ into the z-plane. Then, one can rewrite x-dynamics in (1b) using $z_{\mathsf{n}}^{\mathsf{a}}$ and the nominal components ψ_{n}, ϕ_{n}, and g_{n} as in

$$\begin{aligned}\dot{x} &= A_\nu x + B_\nu\Big(\psi^{\mathsf{T}} z + \phi^{\mathsf{T}} x + g(u_{\mathsf{c}} + a)\Big) \\ &= A_\nu x + B_\nu\Big(\psi_{\mathsf{n}}^{\mathsf{T}} z_{\mathsf{n}}^{\mathsf{a}} + \phi_{\mathsf{n}}^{\mathsf{T}} x + g_{\mathsf{n}} u_{\mathsf{c}} - g(a^\star - a)\Big)\end{aligned} \tag{14}$$

with a new symbol

$$a^\star := \frac{1}{g}\Big(-\psi^{\mathsf{T}} z + \psi_{\mathsf{n}}^{\mathsf{T}} z_{\mathsf{n}}^{\mathsf{a}} + (\phi_{\mathsf{n}}^{\mathsf{T}} - \phi^{\mathsf{T}}) x + (g_{\mathsf{n}} - g) u_{\mathsf{c}}\Big). \tag{15}$$

We now look at a composite full state $\chi := [z_{\mathsf{n}}^{\mathsf{a}}; x; c]$, which can be viewed as the state of the closed-loop system (1) and (2) with $z_{\mathsf{n}}^{\mathsf{a}}$ participating in instead of z. That is, the systems (1), (2), (12), (14), and (15) are compactly represented as

$$\begin{bmatrix}\dot{z}_u \\ \dot{z}_s \\ \dot{\chi}\end{bmatrix} = \begin{bmatrix} S_u & 0 & \mathsf{A}_{12,u} \\ 0 & S_s & \mathsf{A}_{12,s} \\ 0 & 0 & \mathsf{A}_{\mathsf{n}}\end{bmatrix}\begin{bmatrix} z_u \\ z_s \\ \chi\end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \mathsf{B}_{\mathsf{n}}\end{bmatrix} y_r + \begin{bmatrix} 0 \\ 0 \\ \mathsf{B}_a\end{bmatrix}\tilde{a} \tag{16}$$

where $\tilde{a} := a^\star - a$ with some matrices $\mathsf{A}_{12,u}$, $\mathsf{A}_{12,s}$, and B_a.

For now, we suppose that $a^\star(t)$ is available to the attacker, who then takes its attack policy as

$$a(t) = a^\star(t), \quad \forall t \geq 0. \tag{17}$$

Then, the χ-dynamics under the attack (17) becomes the same as that of the nominal closed-loop system (11) (i.e., the χ_{n}-dynamics). Moreover, because the z_u-subsystem in (16) becomes unobservable from the output y and it is anti-stable, the state $z_u(t)$ diverges in most cases while its divergence never influences the output y. The discussion so far is summarized in the following proposition.

Proposition 2: With the attack (17), the solution of the closed-loop system (16) satisfies the following:

(a) For every $[z_u(0); z_s(0); \chi(0)] \in (\mathcal{Z}_0 \times \mathcal{X}_0) \setminus \mathcal{L}_0^\star$ where $\mathcal{L}_0^\star \subset \mathbb{R}^{(n-\nu)+n+n_{\mathsf{c}}}$ is a Lebesgue measure zero set,

$$\|z_u(t)\| \to \infty \ \text{ as } \ t \to \infty.$$

(b) For the solution $\chi_{\mathsf{n}}(t)$ of (11) initiated at $\chi_{\mathsf{n}}(0) = \chi(0)$,

$$\chi(t) = \chi_{\mathsf{n}}(t), \quad \forall t \geq 0.$$

In particular, there are compact sets $\mathcal{Z}_s \subset \mathbb{R}^{n_s}$ and $\mathcal{X} \subset \mathbb{R}^n$ such that $[z_s(t); \chi(t)] \in \mathcal{Z}_s \times \mathcal{X}$, $\forall t \geq 0$. □

We omit the proof due to the page limit.

Remark 3: The first item of the proposition says that success of the attack (17) depends on the initial condition $[z_u(0); z_s(0); \chi(0)]$. The reason for this is the fact that, even for anti-stable system $\dot{\mathsf{x}} = \mathsf{A}\mathsf{x} + \mathsf{B}\mathsf{u}$ with any bounded input signal $\mathsf{u}(t)$, there always exists at least one initial condition $\mathsf{x}(0)$ from which the solution $x(t)$ is bounded, and that the initial condition is determined by all future information of $\mathsf{u}(t)$, $\forall t \geq 0$ [16]. Proposition 2 also says those initial conditions compose a Lebesgue measure zero set, so that this special case hardly occurs. ■

Now we recall that the attack policy (17) is not realistic yet because of two reasons: (i) knowledge of $z(0)$, when the attack is initiated, is needed in order to set $z_{\mathsf{n}}^{\mathsf{a}}(0)$ as in (13), and (ii) the attack signal $a^\star$ is composed of uncertain parameters and unmeasured states. The first item (i) may not be a big problem if the system is in the steady state so that the value of z is easily guessed (at least approximately), or if the uncertainty is not too large and the attacker can employ a state observer using the information of y and u_{c} before the attack starts. Or, in some cases, the information of z may be actually available to the attacker. On the other hand, the second item (ii) looks more challenging. Yet interestingly, this situation is quite familiar in the perspective of robust control. Indeed, if one regards $a^\star$ in (14) as the so-called *lumped disturbance* (or *total disturbance*) [12], [14], then the problem under consideration is converted into how to design a robust controller that estimates and compensates the disturbance $a^\star$. As one of such robust controllers, we will construct a *disturbance observer* (DOB) in the next subsection, which will become our robust attack generator (10) that implements the ideal policy (17) in a practical sense.

C. Robust Zero-dynamics Attack: Implementation of $a = a^\star$ via Disturbance Observer Technique

For the design of DOB, some sets and matrices are defined below. First, take compact sets $\hat{\mathcal{Z}}_s$ and $\hat{\mathcal{X}}$ that strictly contain $\mathcal{Z}_s$ and $\mathcal{X}$ in Proposition 2, respectively. Next, for a given number $\overline{z}_u > 0$, consider the set

$$\mathcal{A}(\overline{z}_u) := \left\{ a^\star \text{ in (15)} : \ \|z_u\| \leq \overline{z}_u, \ [z_s; \chi] \in \hat{\mathcal{Z}}_s \times \hat{\mathcal{X}} \right\}$$

containing all the possible values of $a^\star$ under the variations of g, ψ, and ϕ. This set is clearly bounded, because all the variations of the uncertain parameters are bounded. Finding the set $\mathcal{A}(\overline{z}_u)$ may be a difficult task, and so, let us choose its upper estimate $\hat{\mathcal{A}}(\overline{z}_u)$ that is compact and strictly contains $\mathcal{A}(\overline{z}_u)$. Using this compact set, design a saturation function $\bar{s} : \mathbb{R} \to \mathbb{R}$ that is C^1, bounded, and satisfies

$$\bar{s}(\hat{a}) = \hat{a}, \ \forall \hat{a} \in \hat{\mathcal{A}}(\overline{z}_u) \ \text{ and } \ 0 \leq \frac{\partial \bar{s}}{\partial \hat{a}}(\hat{a}) \leq 1, \ \forall \hat{a} \in \mathbb{R}.$$

Here $\bar{s}$ can be any smooth bounded function whose slope is limited by 1, and is identity on the set $\hat{\mathcal{A}}(\overline{z}_u)$.

In addition, we define a matrix $\Gamma(\tau) := \mathrm{diag}(\tau, \ldots, \tau^\nu) \in \mathbb{R}^{\nu \times \nu}$ for a positive constant τ, and also define vectors $\overline{\phi}_{\mathsf{n}} := [\phi_{\mathsf{n},\nu}; \cdots; \phi_{\mathsf{n},1}] \in \mathbb{R}^\nu$ (with $\phi_{\mathsf{n}} =: [\phi_{\mathsf{n},1}; \ldots; \phi_{\mathsf{n},\nu}]$), and $\alpha := [\alpha_{\nu-1}; \cdots; \alpha_0] \in \mathbb{R}^\nu$. Here, the components α_i, $i = 0, \ldots, \nu - 1$, are selected such that the transfer function

$$W(s) := \frac{s^\nu + \alpha_{\nu-1} s^{\nu-1} + \cdots + \alpha_1 s + (\overline{g}/g_{\mathsf{n}})\alpha_0}{s^\nu + \alpha_{\nu-1} s^{\nu-1} + \cdots + \alpha_1 s + (\underline{g}/g_{\mathsf{n}})\alpha_0} \tag{18}$$

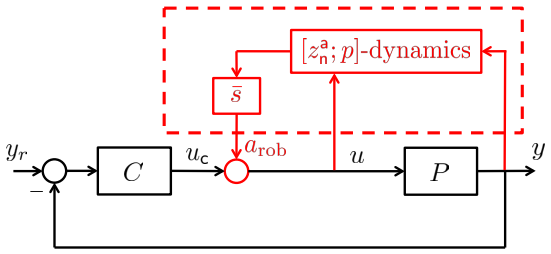


Fig. 2. The proposed attack generator (red dashed block) and the system

is strictly positive real. The design parameters τ and $\overline{z}_u$ used above will be discussed in Theorem 1.

Remark 4: The coefficients α_i that make $W(s)$ strictly positive real can always be obtained by the following two steps. First, select $\alpha_1, \ldots, \alpha_{\nu-1}$ to make $s^{\nu-1}+\alpha_{\nu-1}s^{\nu-2}+\cdots+\alpha_1$ Hurwitz. Then, pick sufficiently small $\alpha_0 > 0$ such that the Nyquist plot of

$$G(s) = \frac{\alpha_0}{s^\nu + \alpha_{\nu-1}s^{\nu-1} + \cdots + \alpha_1 s}$$

does not encircle the disk $D(\underline{g}/g_{\mathsf{n}}, \overline{g}/g_{\mathsf{n}})$ (i.e., the disk in the complex plane whose diameter is the real line segment $[-g_{\mathsf{n}}/\underline{g}, -g_{\mathsf{n}}/\overline{g}]$). For more details, see [13]. ■

Following the DOB structure in [15], a robust attack generator (10) is designed by

$$\dot{p} = \left(A_\nu - \Gamma^{-1}\alpha C_\nu\right) p \tag{19a}$$
$$+ \frac{\alpha_0}{\tau^\nu} B_\nu \left(u_{\mathsf{c}} + a_{\rm rob} + \frac{1}{g_{\mathsf{n}}}\psi_{\mathsf{n}}^{\mathsf{T}} z_{\mathsf{n}}^{\mathsf{a}}\right) + \frac{\alpha_0}{\tau^\nu}\frac{1}{g_{\mathsf{n}}}\left(\overline{\phi}_{\mathsf{n}} + \Gamma^{-1}\alpha\right) y,$$
$$a_{\rm rob} = \bar{s}\left(C_\nu p - \frac{\alpha_0}{\tau^\nu}\frac{1}{g_{\mathsf{n}}} y\right) \tag{19b}$$

together with the $z_{\mathsf{n}}^{\mathsf{a}}$-dynamics (12), where $p \in \mathbb{R}^\nu$ and $z_{\mathsf{n}}^{\mathsf{a}} \in \mathbb{R}^{n-\nu}$ are the states, u_{c} and y are the inputs, and $a_{\rm rob} \in \mathbb{R}$ is the output. (See Fig. 2.) The initial condition $p(0)$ can be anything (as long as it belongs to a compact set) but is set to be zero for convenience, and $z_{\mathsf{n}}^{\mathsf{a}}(0)$ is initiated as in (13).

Briefly explaining how the robust attack generator (12) and (19) works, the following lemma shows that the attack $a_{\rm rob}(t)$ plays the role as an (approximate) estimate of $a^\star(t)$.

Lemma 1: Suppose that $a = a_{\rm rob}$ and

$$\|z_u(t)\| \le \overline{z}_u, \quad [z_s(t); \chi(t)] \in \hat{\mathcal{Z}}_s \times \hat{\mathcal{X}}.$$

Then there exists $\overline{\tau}_0 > 0$ such that for every $\tau \in (0, \overline{\tau}_0)$,

$$\|\tilde{a}(t)\| = \|a_{\rm rob}(t) - a^\star(t)\| \le \frac{k_1}{\tau^\nu} e^{-\lambda_1 (t/\tau)} + k_2 \tau \tag{20}$$

where k_1, k_2, and λ_1 are positive constants independent on τ. □

It is seen from (20) that the smaller τ is, the faster and the more accurate the approximation $a_{\rm rob}(t) \approx a^\star(t)$ is. Consequently, with small τ and large $\overline{z}_u$, the proposed attack generator (12) and (19) can almost recover the attack performance of the ideal attack policy (17), as seen in the following theorem.

Theorem 1: Suppose all assumptions hold. Then, for given $\underline{z}_u > 0$ and $\epsilon > 0$, there exist $\overline{\tau} > 0$ and $\overline{z}_u > \underline{z}_u$ such that for each $[z_u(0); z_s(0); \chi(0)] \in (\mathcal{Z}_0 \times \mathcal{X}_0) \setminus \hat{\mathcal{L}}_0^\star$ where $\hat{\mathcal{L}}_0^\star$ is a Lebesgue measure zero set, the closed-loop system (1), (2), (12), (13), and (19) with $a = a_{\rm rob}$ and $\tau \in (0, \overline{\tau})$ satisfies the following:

(a) There exists a constant $T > 0$ such that

$$\|z_u(T)\| > \underline{z}_u. \tag{21}$$

(b) For $0 \le t \le T$,

$$\|\chi(t) - \chi_{\mathsf{n}}(t)\| < \epsilon \tag{22}$$

where $\chi_{\mathsf{n}}(t)$ is the solution of (11) from $\chi_{\mathsf{n}}(0) = \chi(0)$.

The proofs of Lemma 1 and Theorem 1 are omitted due to the page limit.

IV. Simulation: Power Generating Systems

We consider the scenario that a malicious attack enters a power generating system having a hydro turbine [17], [18]. A state-space representation of the plant with the droop characteristics is given by

$$\dot{\zeta}_1 = -(1/T_P)\zeta_1 + (K_P/T_P)(\zeta_2 - 2\zeta_3), \tag{23a}$$
$$\dot{\zeta}_2 = -(2/T_w)\zeta_2 + (6/T_w)\zeta_3, \tag{23b}$$
$$\dot{\zeta}_3 = -(1/T_G)\zeta_3 + (1/T_G)\left(u_{\mathsf{c}} + a - (1/R)\zeta_1\right), \tag{23c}$$

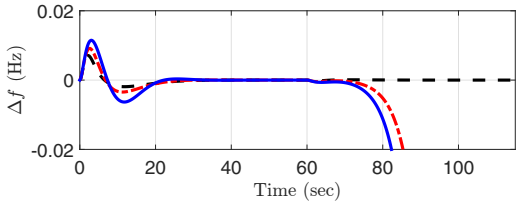
where u_{c} is the input, $y = \zeta_1$ is the output, and $\zeta = [\zeta_1; \zeta_2; \zeta_3] := [\Delta f; \Delta P_G + 2\Delta X_G; \Delta X_G]$ is the state consisting of the incremental frequency deviation Δf (Hz), the change in generator output ΔP_G (p.u.), and the change in governor valve position ΔX_G (p.u.). The constants T_P, T_w, and T_G indicate time constants of load and machine, hydro turbine, and governor, respectively, and R is the speed regulation due to the governor action (Hz/p.u.). The detailed parameters of the plants are given by $K_P = 1$, $T_P = 6$, $T_G = 0.2$, and $R = 0.05$, while $T_w \in [4, 6]$ is uncertain [18]. For robustly stabilizing the uncertain plant (23), a (band-limited) PID-type controller $K(s) = (1.8124s^2 - 18.8558s + 0.1523)/(0.01s^2 + s)$ is designed *a priori*.

The main purpose of the attack is that the valve position ΔX_G increases up to 1.5 p.u., while the frequency derivation Δf and the generating power ΔP_G remain close to those without attack. As a result, the attack leads to overuse of water in a forebay for generating the same amount of power.

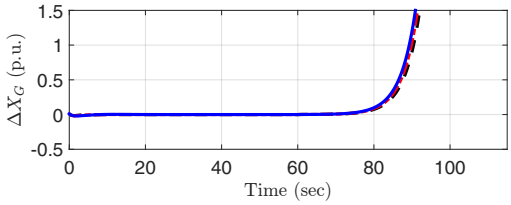
For the attack design, we first represent the hydro-turbine power system (23) in the normal form (1) with a coordinate transformation

$$x_1 := \zeta_1, \; x_2 := -(1/T_P)\zeta_1 + (K_P/T_P)\zeta_2 - (2K_P/T_P)\zeta_3,$$
$$z := \zeta_2 + (3T_P/T_w)(1/K_P)\zeta_1$$

and some constants ϕ_1, ϕ_2, ψ, g, and $S > 0$ (which are possibly uncertain because of T_w). For comparison, let us construct two types of attack generator; one is the conventional zero-dynamics attack (4) for the nominal plant with $T_{w,\mathsf{n}} = 4$; the other is the proposed robust attack (12) and (19) designed with the same $T_{w,\mathsf{n}}$, $\tau = 0.001$, $\overline{z}_u = 1.6$, and a saturation function $\bar{s}(\hat{a})$ whose inactive region is $\hat{\mathcal{A}} = \{\hat{a} : |\hat{a}| \le 1000\}$.

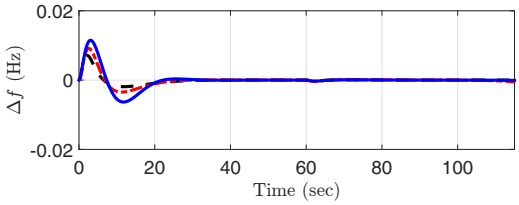


(a) Incremental frequency deviation Δf (Hz)

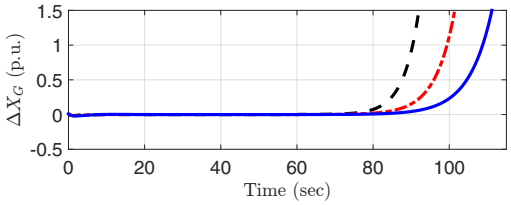


(b) Change in valve position ΔX_G (p.u.)

Fig. 3. Simulation results with the conventional zero-dynamics attack (4) when $T_w = 4 = T_{w,\mathsf{n}}$ (black dashed), $T_w = 5$ (red dash-dotted), and $T_w = 6$ (blue solid)



(a) Incremental frequency deviation Δf (Hz)



(b) Change in valve position ΔX_G (p.u.)

Fig. 4. Simulation results with the proposed robust zero-dynamics attack (12) and (19) when $\tau = 0.001$, $T_w = 4 = T_{w,\mathsf{n}}$ (black dashed), $T_w = 5$ (red dash-dotted), and $T_w = 6$ (blue solid)

Figs. 3 and 4 depict the simulation results of the conventional and the proposed attacks applied to the uncertain plant (23), respectively. These attacks are initiated at $t = 60$ sec when the controlled system is already in the steady state. As shown in these figures, when there is no uncertainty, both attacks work as desired and successfully spoil the plant. However, the conventional scheme (4) immediately fails to be stealthy if it encounters the uncertain plant (Fig. 3), while the proposed attack (12) and (19) remains robust against model uncertainty (Fig. 4). It is also observed from these figures that the modified stealthiness, the item (b) of Problem Statement, is sufficient for the success of the proposed attack. This is because, as discussed in Remark 2, the existing controller $K(s)$ robustly stabilizes the uncertain plant, so all the possible (attack-free) output trajectories $y_0(t)$ (including $y_{\mathsf{n}}(t)$) remain around the zero regardless of model uncertainty, at the moment the attack is initiated.

V. Concluding Remarks

We have shown in this paper that fatal attacks on cyber-physical systems are possible even without exact system knowledge by employing robust control techniques that automatic control community has developed for a long time. Specifically, we have presented a robust zero-dynamics attack that remains stealthy even for uncertain physical systems while forcing the internal states, corresponding to unstable zero dynamics of the system, to diverge. Although in this paper the robust zero dynamics attack is presented for systems with linear dynamics, extending the result to uncertain nonlinear systems seems straightforward. This indicates more research is called for in order to prevent lethal attacks on cyber-physical systems.

References

[1] E. A. Lee, "Cyber physical systems: design challenges," *11th IEEE Symposium on Object Oriented Real-Time Distributed Computing*, pp. 363–369, 2008.

[2] R. Baheti and H. Gill, "Cyber-physical systems," *The Impact of Control Technology,* pp. 161–166, 2011.

[3] S. Gorman, "Electricity grid in U.S. penetrated by spies," *Wall Street Journal*, 2009.

[4] T. Rid, "Cyber war will not take place," *Journal of Strategic Studies*, 2011.

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

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

[7] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Transactions on Automatic Control*, vol. 59, no. 6, pp. 1454–1467, 2014.

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

[9] Y. Mo and B. Sinopoli, "Secure control against replay attacks," *47th Annual Allerton Conference*, pp. 911–918, 2009.

[10] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "Revealing stealthy attacks in control systems," *15th Annual Allerton Conference*, pp. 1806–1813, 2012.

[11] H. K. Khalil, *Nonlinear Systems* (3rd ed.), Prentice Hall, 1996.

[12] H. Shim, G. Park, Y. Joo, J. Back, and N. H. Jo, "Yet another tutorial of disturbance observer: robust stabilization and recovery of nominal performance," to appear at Special Issue, *Control Theory and Technology*, 2016. `http://arxiv.org/abs/1601.02075`

[13] J. Back and H. Shim, "Adding robustness to nominal output-feedback controllers for uncertain nonlinear systems: a nonlinear version of disturbance observer," *Automatica*, vol. 44, no. 10, pp. 2528–2537, 2008.

[14] J. Han, "From PID to active disturbance rejection control," *IEEE Transactions on Industrial Electronics,* vol. 56, no. 3, pp. 900–906, 2009.

[15] J. Back and H. Shim, "Reduced-order implementation of disturbance observers for robust tracking of non-linear systems," *IET Control Theory & Applications*, vol. 8, no. 17, pp. 1940–1948, 2014.

[16] L. R. Hunt, G. Meyer, and R. Su, "Noncausal inverses for linear systems," *IEEE Transactions on Automatic Control*, vol. 41, no. 4, pp. 608–611, 1996.

[17] P. Kundur, *Power System Stability and Control*, McGraw-hill, 1994.

[18] T. Wen, "Unified tuning of PID load frequency controller for power systems via IMC," *IEEE Transactions on Power Systems*, vol. 25, no. 1, pp. 341-350, 2010.