Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

Input Design for Active Dectection of Integrity Attacks using Set-based Approach ⋆

Carlos Trapiello *,** **Vicenç Puig** *,**

* *Research Center for Supervision, Safety and Automatic Control (CS2AC) of the Universitat Politècnica de Catalunya (UPC), Rambla Sant Nebridi 22, 08222, Terrassa, Spain(e-mail: carlos.trapiello@upc.edu).*
** *Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Llorens i Artigas 4-6, 08028 Barcelona, Spain.*

Abstract:
This paper presents the design of an input sequence in order to actively guarantee detectability of integrity attacks. The design of the input sequence is formulated as an optimization problem where the performance degradation imposed in the protected system is minimized while guaranteeing attack detectability by separating the reachable sets of the system in healthy and attacked operation. By considering uncertainties bounded by zonotopes, the design of an optimal open-loop input sequence such that guarantee the separability of the reachable zonotopic sets can be computed by solving a Mixed Integer Quadratic Program (MIQP). Following this approach, attack detection can be guaranteed by: I) forcing a distinct behavior of the system outputs; II) ensuring that residuals under attack will exit the healthy residual set. Furthermore, the present work also considers the imposition of residuals detectability for the specific replay attack scenario affecting an state estimate control system. The effectiveness of the proposals is validated in simulation by means of a numerical example.

Keywords: Attack Detection

1. INTRODUCTION

The introduction of new security vulnerabilities with the development of the complex cyber-physical systems (CPSs), altogether with an increasing number of registered attacks against critical infrastructures (see Sánchez et al. (2019) and the references therein), has drawn attention to the need to develop reliable and secure systems against malicious attacks. In this regard, effective cybersecurity schemes for CPSs must be extended beyond the communications layer, including the study of how attacks affect the estimation and control algorithms in closed-loop systems (Cárdenas et al., 2008).

From an automatic control perspective, deception attacks are specially harmful due to the capability of disrupting the plant operation while remaining undetectable by anomalies detectors, blocking thus the development of possible countermeasures. Throughout the duration of the attack, successful deception attacks must fed system detectors with sets of data consistent with the plant normal operation. This could be done taking advantage of particularities of the system dynamics (zero dynamics attack Teixeira et al. (2012)), or in a broader case, counterfeiting the sensors data. The deception capabilities that are at the core of the previous integrity attacks, has inevitably led to active attack detection schemes such as the so-called physical watermarking techniques used for example in Mo et al. (2015). Working in a probabilistic fashion, these works have to deal with the existing trade-off between detectability rate and the performance degradation imposed in the protected system.

Securing control systems is not a new research field, with the existence of a well-grounded fault detection (FD) and fault tolerant control (FTC) literature (Blanke et al., 2006). Within that field, set-based techniques, which propose a deterministic approach conversely to the most common probabilistic ones, have proved to be a good approach in order to guarantee state estimation and fault diagnosis in systems with uncertainties. Furthermore, several works have treated the set-based active fault detection problem. Among them, it deserves special attention the work of Scott et al. (2014) (further developed in Raimondo et al. (2016)), which, taking advantage of zonotopic set representations, formulates the design of an open-loop optimal sequence of inputs, such that guarantees fault diagnosis among several models, as an MIQP optimization.

According to the aforementioned, the main contribution of this paper is the formulation of an optimal (in the minimum performance degradation sense) input sequence, such that guarantees attack detectability against integrity attacks counterfeiting all the system outputs. This is done under the assumption of discrete linear systems subject to unknown but zonotopically bounded disturbances. The

⋆ This work has been partially funded by AGAUR ACCIO RIS3CAT UTILITIES 4.0 – P7 SECUTIL.

input design considers forcing the attack detection in: I) the system outputs; II) the system residuals. Furthermore, the specific case of an integrity attack against a state estimate control system is also taken into consideration.

The remaining of the paper is structured as follows: the present Section 1 is concluded recalling some basic concepts regarding zonotopic sets. Section 2 presents the problem statement and the considered optimization criterion. In Section 3, the optimization constraints that guarantee attack detectability in the outputs are detailed. Section 4 presents the constraints that ensure attack detectability in the residuals. Section 5 exemplifies the paper proposals by means of numerical simulations. Finally, the main conclusions are drawn in Section 6.

Zonotopes and operations

Let us briefly recall some basic concepts regarding zonotopes. Given a vector $c \in \mathbb{R}^n$ and a set of vectors $\mathcal{H} = \{h_1, ..., h_m\} \subset \mathbb{R}^n$, with $m \geq n$, the generator representation of a zonotope $\mathcal{Z}$ is defined as follows (Le et al. (2013))

$$\mathcal{Z} = \{x \in \mathbb{R}^n : x = c + \sum_{i=1}^{m} \alpha_i h_i; -1 \leq \alpha_i \leq 1\} \quad (1)$$

where c is the center of the zonotope and $H = [h_1, ..., h_m]$ the generator matrix. Hereinafter zonotopes will be denoted as $\mathcal{Z} = \langle c, H \rangle$. Besides, given matrix $L \in \mathbb{R}^{l \times n}$ and the zonotopic sets $\mathcal{Z} = \langle c_z, H_z \rangle$ and $\mathcal{X} = \langle c_x, H_x \rangle$, the sets resulting from linear mappings and Minkowski sum operations are also zonotopes computed as

$$L\mathcal{Z} = \langle Lc_z, LH_z \rangle \quad (2)$$
$$\mathcal{Z} \oplus \mathcal{X} = \langle c_z + c_x, [H_z \;\; H_x] \rangle \quad (3)$$

2. PROBLEM FORMULATION

Let us consider a linear discrete-time system with time k, state vector $x_k \in \mathbb{R}^{n_x}$, input signal $u_k \in \mathbb{R}^{n_u}$, output $y_k \in \mathbb{R}^{n_y}$, process disturbance $w_k \in \mathbb{R}^{n_w}$ and sensor noise $v_k \in \mathbb{R}^{n_v}$. Furthermore, let us differentiate between the system modes $i \in \mathcal{I} \equiv \{h, a\}$, where (h) denotes the *healthy* operation and (a) the system operation under an integrity *attack*. The state space representation of the system model is

$$\begin{aligned} x^i_{k+1} &= Ax^i_k + Bu_k + E_w w^i_k \\ y^i_k &= Cx^i_k + E_v v^i_k \end{aligned} \quad (4)$$

Moreover, the so-called *virtual system* is defined as

$$\begin{aligned} x^v_{k+1} &= Ax^v_k + E_w w^v_k \\ y^v_k &= Cx^v_k + E_v v^v_k \end{aligned} \quad (5)$$

The system initial state, i.e. at $k = 0$, is restricted to $x^i_0 \in \mathcal{X}^i_0$ and $x^v_0 \in \mathcal{X}^v_0$, while the uncertainties satisfy $(w_k, v_k) \in \mathcal{W} \times \mathcal{V}$, $\forall k \in \mathbb{N}$, $\forall i \in \mathcal{I}$. Besides, it is assumed that $\mathcal{X}^i_0$, $\mathcal{X}^v_0$, $\mathcal{W}$ and $\mathcal{V}$ are zonotopic sets with the form

$$\begin{aligned} \mathcal{X}^i_0 &= \langle c^i_{x_0}, H^i_{x_0} \rangle \quad & \mathcal{X}^v_0 &= \langle c^v_{x_0}, H^v_{x_0} \rangle \\ \mathcal{W} &= \langle c_w, H_w \rangle & \mathcal{V} &= \langle c_v, H_v \rangle \end{aligned} \quad (6)$$

Regarding system modes $i \in \mathcal{I}$, the following differences are taken into consideration: the system is considered to be in *healthy* mode (h) if no attacks are being launched against the plant; the system is in *attacked* (a) mode whenever all the system outputs are being substituted by a consistent set of measurements. Conversely, the *virtual system* (5) represents the model that generates the set of replaced outputs

$$\mathcal{Y} = \{y^v_k : k \in [k_s,\ k_f]\} \quad (7)$$

with k_s and k_f being the attack start and final samples, respectively. In the sequel, it is assumed that the signal $u_k \in \mathbb{R}^{n_u}$, injected in order to elucidate the system state i, is unknown to the attacker (u_k is not present in (5)).

Therefore, given a time interval $[0,\ N]$ the goal is to compute an open-loop input sequence $\tilde{u}_{0:N} = (u_0, ..., u_{N-1})$ that guarantees attack detectability, i.e. to deterministically force a distinct behavior between the *healthy* and *attacked* modes. This must be done subject to the polytopic constraint $u_k \in \mathcal{U}$ $(\forall k \in \mathbb{N})$ and under the performance criterion that the injection of the signal has minimum impact with respect the nominal behavior of the plant.

2.1 Set notations

This section briefly presents the computation of the zonotopic representation of the reachable state and output sets for systems with the structure of (4). For a more detailed explanation of similar sets computation the reader is referred to Scott et al. (2014).

Given the sequences $\tilde{u}_{0:k} = (u_0, ..., u_{k-1}) \in \mathbb{R}^{kn_u}$ and $\tilde{w}^i_{0:k} \in \mathbb{R}^{kn_w}$ (with $\tilde{w}^i_{0:k}$ defined similarly), then, the functions $\phi^i_k(\tilde{u}_{0:k}, x^i_0, \tilde{w}^i_{0:k})$ and $\psi^i_k(\tilde{u}_{0:k}, x^i_0, \tilde{w}^i_{0:k}, v^i_k)$ define the state and output of system (4) at time k. For some $l, k \in \mathbb{N}$ such that $0 \leq l \leq k \leq N$, let us define the sequences $\tilde{\phi}^i_{l:k}(\tilde{u}, x^i_0, \tilde{w}^i) = (\phi^i_l, ..., \phi^i_k)$ and $\tilde{\psi}^i_{l:k}(\tilde{u}, x^i_0, \tilde{w}^i, \tilde{v}^i) = (\psi^i_l, ..., \psi^i_k)$ (where the dependences of ϕ_j and ψ_j $\forall j \in [l,\ k]$ were omitted for simplicity). Taking into consideration the uncertainties in the initial state and process/noise disturbances, the reachable state and output sets on the interval $[l,\ k]$ are defined as

$$\begin{aligned} \tilde{\Phi}^i_{l:k}(\tilde{u}) &\equiv \{\tilde{\phi}^i_{l:k}(\tilde{u}, x^i_0, \tilde{w}^i) : (x^i_0, \tilde{w}^i) \in \mathcal{X}^i_0 \times \tilde{\mathcal{W}}\} \\ \tilde{\Psi}^i_{l:k}(\tilde{u}) &\equiv \{\tilde{\psi}^i_{l:k}(\tilde{u}, x^i_0, \tilde{w}^i, \tilde{v}^i) : (x^i_0, \tilde{w}^i, \tilde{v}^i) \in \mathcal{X}^i_0 \times \tilde{\mathcal{W}} \times \tilde{\mathcal{V}}\} \end{aligned} \quad (8)$$

where $\tilde{\mathcal{W}} = \mathcal{W} \times ... \times \mathcal{W}$ and $\tilde{\mathcal{V}} = \mathcal{V} \times ... \times \mathcal{V}$ with k and $k+1$ products, respectively.

Hence, the propagation of model (4) recursively define the extended matrices $\tilde{A}$, $\tilde{B}$, $\tilde{C}$, $\tilde{E}_w$, $\tilde{E}_v$ such that

$$\begin{aligned} \tilde{\phi}^i_{l:k}(\tilde{u}, x^i_0, \tilde{w}^i) &= \tilde{A}x^i_0 + \tilde{B}\tilde{u} + \tilde{E}_w\tilde{w}^i \\ \tilde{\psi}^i_{l:k}(\tilde{u}, x^i_0, \tilde{w}^i, \tilde{v}^i) &= \tilde{C}\tilde{\phi}^i_{l:k}(\tilde{u}, x^i_0, \tilde{w}^i) + \tilde{E}v\tilde{v}^i \end{aligned} \quad (9)$$

where the reachable state and output sets are computed as

$$\begin{aligned} \tilde{\Phi}^i_{l:k}(\tilde{u}) &= \tilde{A}\mathcal{X}^i_0 \oplus \tilde{B}\tilde{u} \oplus \tilde{E}_w\tilde{\mathcal{W}} \\ \tilde{\Psi}^i_{l:k}(\tilde{u}) &= \tilde{C}\tilde{\Phi}^i_{l:k}(\tilde{u}) \oplus \tilde{E}_v\tilde{\mathcal{V}} \end{aligned} \quad (10)$$

Under the zonotopic assumptions expressed in (6), the resulting sets in (10) are also zonotopes denoted as

$$\begin{aligned} \tilde{\Phi}^i_{l:k}(\tilde{u}) &= \langle c^{\phi^i}_{l:k}(\tilde{u}), H^{\phi^i}_{l:k} \rangle \\ \tilde{\Psi}^i_{l:k}(\tilde{u}) &= \langle c^{\psi^i}_{l:k}(\tilde{u}), H^{\psi^i}_{l:k} \rangle \end{aligned} \quad (11)$$

being the generator matrices and the centers

$$\begin{aligned} c^{\phi^i}_{l:k}(\tilde{u}) &= \tilde{\phi}^i{}_{l:k}(\tilde{u}, c^i_0, c_{\tilde{w}}) \qquad & H^{\phi^i}_{l:k} &= [\tilde{A}H^i_{x_0} \;\; \tilde{E}_w H_{\tilde{w}}] \\ c^{\psi^i}_{l:k}(\tilde{u}) &= \tilde{\psi}^i{}_{l:k}(\tilde{u}, c^i_0, c_{\tilde{w}}, c_{\tilde{v}}) \qquad & H^{\psi^i}_{l:k} &= [\tilde{C}H^{\phi^i}_{l:k} \;\; \tilde{E}_v H_{\tilde{v}}] \end{aligned} \tag{12}$$

with $c_{\tilde{w}} = (c_w, ..., c_w)$ and block-diagonal matrix $H_{\tilde{w}} = diag(H_w, ..., H_w)$ ($c_{\tilde{v}}$, $H_{\tilde{v}}$ are defined similarly).

Note that the injected sequence $\tilde{u}$ only affects the displacement of the set centers, and not the set size. In this regard, let us rewrite the centers expression at time k, by differentiating the affine terms with the injected sequence $\tilde{u}_{0:k}$ as

$$\begin{aligned} c^{\phi^i}_k(\tilde{u}) &= c^{\phi^i}_k(0) + B^{\phi}_k \tilde{u}_{0:k} \\ c^{\psi^i}_k(\tilde{u}) &= c^{\psi^i}_k(0) + B^{\psi}_k \tilde{u}_{0:k} \end{aligned} \tag{13}$$

with

$$B^{\phi}_{k+1} = [AB^{\phi}_k \;\; B] \qquad B^{\psi}_k = CB^{\phi}_k \tag{14}$$

Taking into consideration the extended system matrices presented in (10), then, the centers of the reachable state and output sets given an interval $[l,\ k]$ can be expressed as

$$\begin{aligned} c^{\phi^i}_{l:k}(\tilde{u}) &= c^{\phi^i}_{l:k}(0) + \tilde{B}\tilde{u} \\ c^{\psi^i}_{l:k}(\tilde{u}) &= c^{\psi^i}_{l:k}(0) + \tilde{C}\tilde{B}\tilde{u} \end{aligned} \tag{15}$$

2.2 Performance objective

It is desired that the signal injected in the time interval $[0, N]$ has minimum impact on the protected system performance. In this regard, the imposed performance degradation with respect the nominal system (no external signal is injected), is characterized by means of the center of the state reachability set which, according to (13), evolves as $c^{\phi^i}_k(\tilde{u})$ for the protected system and as $c^{\phi^i}_k(0)$ for the nominal system. For both systems starting in the same initial state x_0 ($\delta c^{\phi}_0 = 0$), the relative displacement of the center at time $k > 0$ is defined as

$$\delta c^{\phi}_k = \begin{cases} B^{\phi}_k \tilde{u}_{0:k} = \sum_{l=0}^{k-1} A^{k-1-l} B u_l & \text{if } 0 < k \leq N \\ A^{k-N} B^{\phi}_N \tilde{u}_{0:N} & \text{otherwise} \end{cases} \tag{16}$$

and the cost function weighting the effect of the sequence $\tilde{u}_{0:N}$ in the performance degradation is formulated as

$$J(\tilde{u}) = \sum_{j=1}^{\infty} (\delta c^{\phi}_j)^T R \delta c^{\phi}_j \tag{17}$$

with R positive semidefinite.

Note that from a practical point of view, the effect of the signal $\tilde{u}_{0:N}$ vanishes several samples after its injection, where the number of samples depends on the the system dynamics. Thus, in order to bound the extension of the cost function (17), let us define an $\epsilon > 0$, $\epsilon \in \mathbb{R}$ arbitrarily small. According to (16), at $k = N + s$, the effect of the last term of the input sequence, i.e. u_{N-1}, in δc^{ϕ}_k is given by $A^s B u_{N-1}$. Therefore, the effect of $\tilde{u}_{0:N}$ is considered negligible if

$$||A^s B||_\infty \leq ||A^s||_\infty \cdot ||B||_\infty < \epsilon \tag{18}$$

Diagonalizing A with respect its eigenvector matrix T, i.e. $A = T\Lambda T^{-1}$, and expressing the spectral radius of A as $\rho(A) = ||\Lambda||_\infty$, it follows

$$||A^s||_\infty = ||T\Lambda^s T^{-1}||_\infty \leq \kappa \rho(A)^s \quad \forall s \in \mathbb{N} \tag{19}$$

where κ is the condition number of A with respect to its eigenproblem: $\kappa = ||T||_\infty \cdot ||T^{-1}||_\infty$. Denoting $\omega = ||B||_\infty$, and replacing (19) in (18), the effect of the last term of the sequence $\tilde{u}_{0:N}$ (and therefore all the others) can be considered negligible for $s \in \mathbb{N}^+$ such that fulfills

$$s > \frac{log(\epsilon) - log(\kappa\omega)}{log(\rho(A))} \tag{20}$$

And therefore the considered cost function is

$$J_{N+s}(\tilde{u}) = \sum_{j=1}^{N+s} (\delta c^{\phi}_j)^T R \delta c^{\phi}_j \tag{21}$$

The necessary conditions that $\tilde{u}_{0:N}$ must met in order to guarantee attack detectability in the system outputs and in the residuals are presented in Section 3 and Section 4, respectively.

3. DETECTION IN THE OUTPUTS

The substitution of the real system outputs by the consistent set $\mathcal{Y}$ implies that, after the injection of the external signal $\tilde{u}_{0:N}$, the system outputs will behave as if the signal was not introduced, i.e. forcing an effect similar to an output fault. This fact entails that the designed sequence $\tilde{u}_{0:N}$ must accomplish the separation of the *healthy* output reachable set $\tilde{\Psi}^H_{0:N}(\tilde{u})$ with respect the output reachable set when no input is injected $\tilde{\Psi}_{0:N}(0)$. Below, the worst separability case is considered, that is, when the real output and the substituted output reachable sets share the same initial set $\mathcal{X}_0$.

According to the problem statement, after the injection of the signal $\tilde{u}_{0:N}$, the sequence of registered outputs $\tilde{y}_{0:N} = (y_0, ..., y_N)$ will exist in

$$\tilde{y}_{0:N} \in \tilde{\Psi}^H_{0:N}(\tilde{u}) \cup \tilde{\Psi}_{0:N}(0) \tag{22}$$

Thus, in order to guarantee the attack detectability by analyzing $\tilde{y}_{0:N}$, the following condition must hold

$$\tilde{\Psi}^H_{0:N}(\tilde{u}) \cap \tilde{\Psi}_{0:N}(0) = \emptyset \tag{23}$$

Similar to Raimondo et al. (2016) let us denote as $\mathcal{L}_N$ the set of all input sequences such that fulfill (23). Thus, an optimal $\tilde{u}_{0:N}$ is defined as a solution of

$$inf\{J(\tilde{u}) : \tilde{u}_{0:N} \in \tilde{U}_N \cap \mathcal{L}_N\} \tag{24}$$

where the fact that $\mathcal{L}_N$ is open, implies the imposition of ϵ-solutions of the infimum. Besides, note the complexity of solving (24) due to the non convexity of $\mathcal{L}_N$.

Following the proposal of Scott et al. (2014), an optimization problem like (24) can be reformulated as an MIQP by means of expressing the set of separating inputs $\mathcal{L}_N$ as

$$\mathcal{L}_N = \{\tilde{u}_{0:N} : \Xi \tilde{u}_{0:N} \notin Z\} \tag{25}$$

where according to (15): $\Xi = \tilde{C}\tilde{B}$ and

$$Z = \{[H^{\Psi}_{0:N} - H^{\Psi}_{0:N}], c^{\psi}_{0:N}(0) - c^{\psi}_{0:N}(0)\} = \{[H^{\Psi}_{0:N} - H^{\Psi}_{0:N}], 0\} \tag{26}$$

The imposition of $\Xi\tilde{u}_{0:N}$ being outside the set $\mathcal{Z}$ can be easily formulated as a linear program that turns optimization (24) into a bilevel program. Finally, this bilevel program can be reformulated as an MIQP by replacing the inner linear program by the corresponding necessary and sufficient conditions of optimality.

Remark 1: Polytopic set $\mathcal{U}$ can be used for imposing the injection of the designed sequence through a subset of secure input channels. Defining thus a new B' input matrix.

Remark 2: Given the matrices A, B', C the output controllability matrix $\mathcal{OC}$ must have $rank(\mathcal{OC}) > 0$ in order to guarantee the existence of a solution in a finite time horizon N. This follows straightforwardly from the structure of Ξ in (25).

4. DETECTION IN THE RESIDUALS

Forcing attack detection in the system outputs entails the construction of highly dimensional zonotopes, as well as certain knowledge regarding the system state at the start of the injection sequence. In order to avoid these drawbacks, the present section considers that attack detectability is forced by guaranteeing that the *attacked* residuals at time $k = N$, will exit the *healthy* reachable residual set at that time.

Let us consider that the residual generation is based on a standard Luenberger observer with the structure

$$\begin{aligned} \hat{x}_{k+1} &= A\hat{x}_k + Bu_k + L(y_k - \hat{y}_k) \\ \hat{y}_k &= C\hat{x}_k \end{aligned} \tag{27}$$

where $\hat{x}_0 \in \hat{\mathcal{X}}_0 = \langle c_{x_0}, 0\rangle$ and the online residual generation is performed as

$$r_k = y_k - \hat{y}_k \tag{28}$$

Below two different scenarios are considered: 1) the attack is deployed against a monitoring station that is supervising the system performance; 2) the standard structure in the literature regarding replay attacks (see Mo and Sinopoli (2009)), where the state estimator that generates the residuals is also used to close the feedback control loop.

4.1 Monitoring station

Let us characterize the residual generation for the different modes $i \in \mathcal{I}$. The residual generation in *healthy* (h) mode follows

$$r_k^h = y_k^h - \hat{y}_k^h = C(x_k^h - \hat{x}_k^h) + E_v v_k \tag{29}$$

where independently of the injected signal u_k, the dynamics of the state estimation error $e_k^h = x_k^h - \hat{x}_k^h$ are governed by

$$e_{k+1}^h = (A - LC)e_k^h + E_w w_k - LE_v v_k \tag{30}$$

with $e_0^h \in \mathcal{E}_0^h = \mathcal{X}_0^h \oplus \langle -c_{x_0}^h, 0\rangle = \langle 0, H_{x_0}^h\rangle$.

Nevertheless, the residuals in the *attacked* (a) mode follows

$$r_k^a = y_k^v - \hat{y}_k^a = C(x_k^v - \hat{x}_k^a) + E_v v_k \tag{31}$$

Note that (31) is comparing false data: $y_k^v \in \mathcal{Y}$, with the estimations that yield an observer operating also based on false measurements as

$$\begin{aligned} \hat{x}_{k+1}^a &= A\hat{x}_k^a + Bu_k + L(y_k^v - \hat{y}_k^a) \\ \hat{y}_k^a &= C\hat{x}_k^a \end{aligned} \tag{32}$$

Taking into consideration (5) and (32), the new estimation error $e_k^a = x_k^v - \hat{x}_k^a$ is governed by

$$e_{k+1}^a = (A - LC)e_k^a - Bu_k + E_w w_k - LE_v v_k \tag{33}$$

with $e_0^a \in \mathcal{E}_0^a = \mathcal{X}_0^v \oplus \langle -c_{x_0}^a, 0\rangle = \langle c_{x_0}^v - c_{x_0}^a, H_{x_0}^v\rangle$. Note the dependence of this state estimation error with the injected signal u_k.

Denoting as $\sigma_k^i(\tilde{u})$ the residual reachable set of mode i at time k, attack detectability is guaranteed in a maximum of N steps if

$$\sigma_N^H(\tilde{u}) \cap \sigma_N^A(\tilde{u}) = \emptyset \tag{34}$$

In order to build the previous sets, let us define the vectors $e_k^\star = [e_k^h, e_k^a]^T$, $w_k^\star = [w_k, w_k]^T$, $v_k^\star = [v_k, v_k]^T$ and $r_k^\star = [r_k^h, r_k^a]^T$, and gather (29) to (33) in the formulation

$$\begin{aligned} e_{k+1}^\star &= A^\star e_k^\star + B^\star u_k + E_w^\star w_k^\star + D^\star v_k^\star \\ r_k^\star &= C^\star e_k^\star + E_v^\star v_k^\star \end{aligned} \tag{35}$$

with

$$\begin{aligned} A^\star &= \begin{bmatrix} A - LC & 0 \\ 0 & A - LC \end{bmatrix} \quad B^\star = \begin{bmatrix} 0 \\ -B \end{bmatrix} \quad C^\star = \begin{bmatrix} C & 0 \\ 0 & C \end{bmatrix} \\ E_w^\star &= \begin{bmatrix} E_w & 0 \\ 0 & E_w \end{bmatrix} \quad D^\star = \begin{bmatrix} -LE_v & 0 \\ 0 & -LE_v \end{bmatrix} \quad E_v^\star = \begin{bmatrix} E_v & 0 \\ 0 & E_v \end{bmatrix} \end{aligned} \tag{36}$$

with both systems initialized for the worst detectability case, i.e. with the integrity attack perfectly deceiving the anomalies detector: $\mathcal{E}_0^a = \mathcal{E}_0^h$.

Following the formulation presented in Section 2.1, and denoting as $\tilde{A}^\star$, $\tilde{B}^\star$,$\tilde{C}^\star$, $\tilde{E_w}^\star$,$\tilde{E_v}^\star$, $\tilde{D}^\star$ the extended matrices similar to (9), the reachable set at time k of the residuals $r_k^\star$ is the zonotope: $\Gamma_k(\tilde{u}) = \langle c_k^r(\tilde{u}), H_k^r\rangle$, where the effect of the injected signal in the center displacement can be separated as

$$c_k^r(\tilde{u}) = c_k^r(0) + \tilde{C}_k^\star \tilde{B}_k^\star \tilde{u} \tag{37}$$

with $c_k^r(0)$ being the output of (35) when $x_0 = c_{x_0}$, $(w_k, v_k) = (c_w, c_v)$ and $u_k = 0,\ \forall k$.

Therefore, considering the gathered construction of the residuals reachable sets, condition (34) is reformulated as

$$[I \ -I]\Gamma_N(\tilde{u}) = \emptyset \tag{38}$$

where Applying Lemma 2 in Scott et al. (2014) the set of separating inputs Ω_N is such that

$$\Omega_N = \{\tilde{u}_{0:N} : [I \ -I]\Xi_N^\star \tilde{u}_{0:N} \notin [I \ -I]\Gamma_N(0)\} \tag{39}$$

with $\Gamma_N(0) = \langle c_N^r(0), H_H^r\rangle$ and $\Xi_N^\star = \tilde{C}_N^\star \tilde{B}_N^\star$.

Similar to the procedure presented in Section 3, the computation of

$$inf\{J(\tilde{u}) : \tilde{u}_{0:N} \in \tilde{U}_N \cap \Omega_N\} \tag{40}$$

can be efficiently solved by reformulating it as an MIQP problem.

4.2 Replay attack on a state estimate feedback controlled system

Replay attacks constitute a specific case of integrity attacks, where the substituted set $\mathcal{Y}$ is obtained by recording previous measurements during the stationary. Under such steady-state conditions, replay attacks launched against state estimate feedback control systems are able to deceive anomalies detectors under certain conditions regarding the controller and observer gains (see Mo and Sinopoli (2009)). Consequently, let us reformulate the development carried out in Section (4.1), by taking into consideration that the plant is controlled by a state estimation feedback control law of the form

$$u_{cl} = -K\hat{x} \tag{41}$$

For this case, residual generation in *healthy* mode will follow (29)-(30). However, residuals in the *attacked* mode can be rewritten as

$$r_k^a = y_k^v - \hat{y}_k^a = y_k^v - \hat{y}_k^v + (\hat{y}_k^v - \hat{y}_k^a) = r_k^v + C(\hat{x}_k^v - \hat{x}_k^a) \quad (42)$$

where the dynamics of r_k^v are the same than (29)-(30). On the other hand, taking into consideration the equivalent estimations of

$$\begin{aligned} \hat{x}_{k+1}^v &= (A - BK)\hat{x}_k^v + L(y_k^v - \hat{y}_k^v) \\ \hat{y}_k^v &= Cx_k^v \end{aligned} \quad (43)$$

and the estimations of the state observer under attack

$$\begin{aligned} \hat{x}_{k+1}^a &= (A - BK)\hat{x}_k^a + Bu_k + L(y_k^v - \hat{y}_k^a) \\ \hat{y}_k^a &= Cx_k^a \end{aligned} \quad (44)$$

Then, the dynamics of $\hat{e}_k = \hat{x}_k^v - \hat{x}_k^a$ are governed by

$$\hat{e}_{k+1} = (A - BK - LC)\hat{e}_k - Bu_k \quad (45)$$

with $\hat{e}_0 = c_{x_0}^v - c_{x_0}^a$, being the worst detectability case with $r_k^v = r_k^h$ and $\hat{e}_0 = 0$. Note that if $(A - BK - LC)$ is not a Schur stable, i.e. its eigenvalues are outside the unit circle, the attack detection in (42) is guaranteed, however the system will unstabilize through (41).

In order to impose the detectability condition (34), let us build the residuals reachable set of the gathered system (denoting the vector $e_k^+ = [e_k^h,\, e_k^v,\, \hat{e}_k]^T$)

$$\begin{aligned} e_{k+1}^+ &= A^+ e_k^+ + B^+ u_k + E_w^+ w_k^\star + D^+ v_k^\star \\ r_k^\star &= C^+ e_k^+ + E_v^\star v_k^\star \end{aligned} \quad (46)$$

where

$$A^+ = \begin{bmatrix} (A-LC) & 0 & 0 \\ 0 & (A-LC) & 0 \\ 0 & 0 & (A-BK-LC) \end{bmatrix} \quad B^+ = \begin{bmatrix} 0 \\ 0 \\ -B \end{bmatrix}$$
$$E_w^+ = \begin{bmatrix} E_w & 0 \\ 0 & E_w \\ 0 & 0 \end{bmatrix} \quad D^+ = \begin{bmatrix} -LE_v & 0 \\ 0 & -LE_v \\ 0 & 0 \end{bmatrix} \quad C^+ = \begin{bmatrix} C & 0 & 0 \\ 0 & C & C \end{bmatrix} \quad (47)$$

and worst case initial condition set $\{\mathcal{E}_0^h\} \times \{\mathcal{E}_0^h\} \times \{0\}$.

Once the $r_N^\star$ reachable set is computed, same development as the presented in Section 4.1 could be employed in order to reformulate the required separability condition as a MIQP problem.

5. NUMERICAL EXAMPLE

Let us consider a discrete time LTI system (see (4)) with the following system matrices

$$A = \begin{bmatrix} 0.6 & 0.3 & -0.2 \\ 0.4 & 0.2 & 0.7 \\ 0.5 & -0.3 & 0.5 \end{bmatrix} \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \quad E_w = \begin{bmatrix} 0.5 & 0 \\ 0.1 & 0.2 \\ 0.1 & 0 \end{bmatrix}$$

initial conditions $x_0 \in \mathcal{X}_0 = \langle [3\ 3\ 3]^T, 0.2I_3 \rangle$ and process disturbances confined in $w_k \in \mathcal{W} = \langle [0\ 0]^T, I_2 \rangle\ \forall k$.

5.1 Detection in the outputs

For representative purposes, let us consider a single output system with matrices

$$C = [1\ 0\ 0] \qquad E_v = 0.2$$

with $v_k \in \mathcal{V} = \langle c_v, H_v \rangle = \langle 0, 1 \rangle\ \forall k$. Furthermore, the control actions are restricted to meet $u_k \in \mathcal{U}\ \forall k$, where

$$\mathcal{U} = \{u = [u_1\ u_2]^T \in \mathbb{R}^2 : |u_1| \leq 2, |u_2| \leq 1\}$$

For $\epsilon = 1e^{-3}$, according to (20), it is considered that the effect of the input sequence has vanished at $k = N + s$, with $s = 50 > 49.4$. Table 1 presents the obtained input sequences $\tilde{u}_{0:N}$ for the time horizons $N = \{1, 2, 3\}$. It can be seen how as the number of degrees of freedom increases (N increases), the performance degradation imposed in the protected system is reduced.

Table 1.

N	u_0	u_1	u_2	J_{N+50}
1	$\begin{bmatrix} +1.8584 \\ -0.8179 \end{bmatrix}$	-	-	12.851
2	$\begin{bmatrix} +1.8584 \\ +0.4415 \end{bmatrix}$	$\begin{bmatrix} -1.4306 \\ -0.7323 \end{bmatrix}$	-	4.821
3	$\begin{bmatrix} -0.9438 \\ -0.0388 \end{bmatrix}$	$\begin{bmatrix} +1.8738 \\ +0.5838 \end{bmatrix}$	$\begin{bmatrix} -1.0114 \\ -0.1912 \end{bmatrix}$	3.060

Fig. 1 depicts the obtained separation of the output reachable zonotopic sets for the *healthy* and *attacked* models for $\tilde{u}_{0:3}$, as well as the recorded output sequences $\tilde{y}_{0:3}$ after 500 simulations with the system in *healthy* mode and 500 in the *attacked* mode. The frame axes of Fig. 1 represent the different time instants (note that y_0 is not depicted since both systems share the output reachable set at $k = 0$).

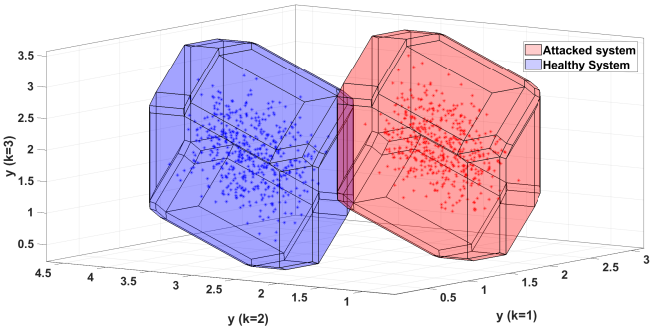

Fig. 1. Detection in the outputs for $\tilde{u}_{0:3}$

5.2 Detection in the residuals

In order to exemplify the proposed detection in the system residuals, let us consider now the matrices

$$C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad E_v = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix}$$

with $v_k \in \mathcal{V} = \langle [0\ 0]^T, I_2 \rangle\ \forall k$, and the set $\mathcal{U}$ defined as

$$\mathcal{U} = \{u = [u_1\ u_2]^T \in \mathbb{R}^2 : |u_1| \leq 1, |u_2| \leq 1\}$$

The observer gain L used is the stationary value such that minimizes the size of the estimation error zonotope (see Combastel (2015))

$$L = \begin{bmatrix} 0.5019 & 0.4777 & 0.4857 \\ 0.2027 & 0.2292 & -0.0749 \end{bmatrix}^T$$

For $N = 5$, the obtained optimal input sequence that forces residual reachable set separation at $k = N$ is

$$\tilde{u}_{0:5} = \left(\begin{bmatrix} -0.7976 \\ 0.2206 \end{bmatrix}, \begin{bmatrix} -0.3661 \\ +1 \end{bmatrix}, \begin{bmatrix} -1 \\ +1 \end{bmatrix}, \begin{bmatrix} 0.9885 \\ +1 \end{bmatrix}, \begin{bmatrix} +1 \\ -1 \end{bmatrix} \right)$$

with a cost function of $J_{55} = 16.1841$.

Fig. 2 plots the obtained temporal evolution for $k \in [0,\ 5]$ of the *healthy* (blue) and *attacked* (red) residual sets. The blue and red clouds of points represent 500 healthy and 500 attacked simulations, respectively. Note that, despite

detection is guaranteed at $k = 5$, the attacked residuals could exit the healthy set at any time before, yielding thus a sooner detection.

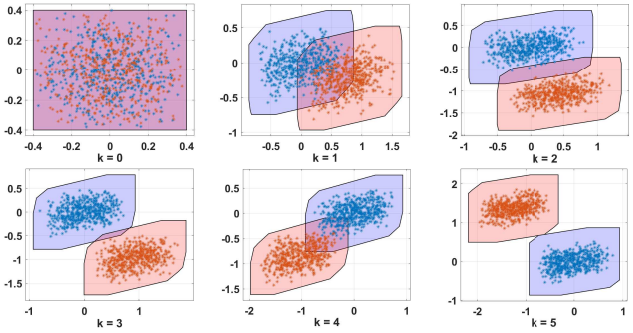

Fig. 2. Residuals evolution $k \in [0\ 5]$

For the different simulations aforementioned, Fig. 3 represents the output sequence $\tilde{y}_{0:5}$ for the protected (blue) and nominal (red) system.

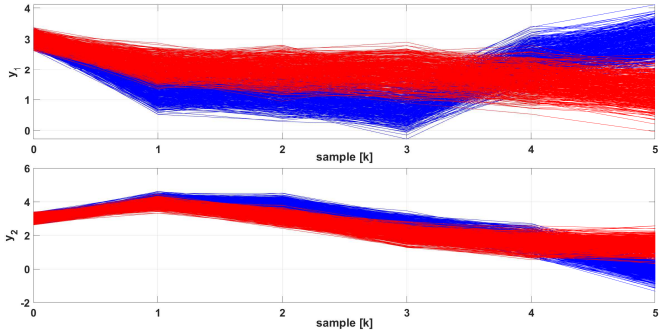

Fig. 3. Output temporal evolution $k \in [0\ 5]$

5.3 Replay attack example

Let us consider a state estimate feedback system with controller gain

$$K = \begin{bmatrix} -0.4224 & 0.4334 & -0.0813 \\ 0.5294 & -0.6776 & 0.631 \end{bmatrix}$$

such that $eig(A - BK - LC) = \{0.8,\ 0.85,\ 0.9\}$

Besides, let us consider a replay attack scenario where an attacker secretly records data during the interval $k \in [100,\ 200]$ and replay it back for $k \in [400,\ 500]$. At $k = 450$, it is desired to inject a $N = 3$ horizon signal in order to elucidate the system mode. Note that at $k = 450$, the initial estimation error $\mathcal{E}_0$ considered in the input design, is the estimation error set at steady-state conditions. The obtained sequence is

$$\tilde{u}_{0:3} = \left(\begin{bmatrix} 0.3488 \\ -0.4466 \end{bmatrix}, \begin{bmatrix} 0.2118 \\ -0.2096 \end{bmatrix}, \begin{bmatrix} -0.0270 \\ -0.7403 \end{bmatrix} \right)$$

with $J_{53} = 1.0842$.

Fig. 4 depicts the stationary *healthy* residual set (blue zonotope) altogether with the registered residuals $\tilde{r}_{450:453}$ for a total of 500 replay attack simulations. It can be seen that in most of the cases detectability is achieved in $k = 452$ (yellow points), and that at $k = 453$ (green points) detectability is guaranteed with all the recorded residuals laying outside of the *healthy* zonotope.

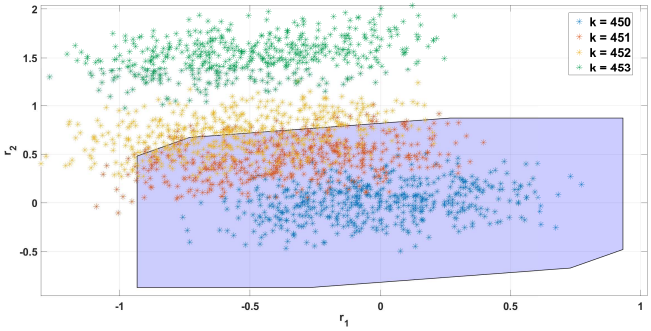

Fig. 4. Replay attack residual space

6. CONCLUSIONS

Despite the inherent conservatism of set-based techniques, their capability to deterministically discuss the state of the system offer an effective approach for security-related issues. Furthermore, the possibility of designing an optimal input sequence that guarantees attack detectability with unitary probability, allows to directly face the existing trade-off between detectability rate and performance degradation existing in physical watermarking strategies. The formulation of the proposed techniques to more complex attack policies as well as studying the effect of the observer gain in the required performance degradation, are identified as future research directions.

REFERENCES

Blanke, M., Kinnaert, M., Lunze, J., Staroswiecki, M., and Schröder, J. (2006). *Diagnosis and fault-tolerant control*, volume 2. Springer.

Cárdenas, A.A., Amin, S., and Sastry, S. (2008). Research challenges for the security of control systems. In *HotSec*.

Combastel, C. (2015). Zonotopes and kalman observers: Gain optimality under distinct uncertainty paradigms and robust convergence. *Automatica*, 55, 265–273.

Le, V.T.H., Stoica, C., Alamo, T., Camacho, E.F., and Dumur, D. (2013). *Zonotopes: From guaranteed state-estimation to control*. John Wiley & Sons.

Mo, Y. and Sinopoli, B. (2009). Secure control against replay attacks. In *47th annual Allerton conference on communication, control, and computing*, 911–918. IEEE.

Mo, Y., Weerakkody, S., and Sinopoli, B. (2015). Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs. *IEEE Control Systems Magazine*, 35, 93–109.

Raimondo, D.M., Marseglia, G.R., Braatz, R.D., and Scott, J.K. (2016). Closed-loop input design for guaranteed fault diagnosis using set-valued observers. *Automatica*, 74, 107–117.

Sánchez, H.S., Rotondo, D., Escobet, T., Puig, V., and Quevedo, J. (2019). Bibliographical review on cyber attacks from a control oriented perspective. *Annual Reviews in Control*.

Scott, J.K., Findeisen, R., Braatz, R.D., and Raimondo, D.M. (2014). Input design for guaranteed fault diagnosis using zonotopes. *Automatica*, 50(6), 1580–1589.

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