

## On reachable sets of hidden CPS sensor attacks

**Citation for published version (APA):**

Murguia, C., & Ruths, J. (2017). On reachable sets of hidden CPS sensor attacks. *arXiv*, Article 1710.06967v1.

**Document status and date:**

Published: 01/10/2017

**Document Version:**

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

**Please check the document version of this publication:**

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

**General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

[www.tue.nl/taverne](http://www.tue.nl/taverne)

**Take down policy**

If you believe that this document breaches copyright please contact us at:

[openaccess@tue.nl](mailto:openaccess@tue.nl)

providing details and we will investigate your claim.

# On Reachable Sets of Hidden CPS Sensor Attacks

Carlos Murguia and Justin Ruths

**Abstract**—For given system dynamics, observer structure, and observer-based fault/attack detection procedure, we provide mathematical tools – in terms of Linear Matrix Inequalities (LMIs) – for computing outer ellipsoidal bounds on the set of estimation errors that attacks can induce while maintaining the alarm rate of the detector equal to its attack-free false alarm rate. We refer to these sets to as *hidden reachable sets*. The obtained ellipsoidal bounds on hidden reachable sets quantify the attacker’s potential impact when it is constrained to stay hidden from the detector. We provide tools for minimizing the volume of these ellipsoidal bounds (minimizing thus the reachable sets) by redesigning the observer gains. Simulation results are presented to illustrate the performance of our tools.

## I. INTRODUCTION

There has recently been significant interest and work in the broad area of security of cyber-physical systems (CPS), see for example [1]-[8]. This topic investigates the properties of conventional control systems in the presence of adversarial disturbances. Control theory has shown great ability to robustly deal with disturbances and uncertainties [9]. However, adversarial attacks raise all-new issues due to the aggressive and strategic nature of the disturbances that attackers might inject into the system.

This paper focuses on attack detection and attack capabilities in CPSs. A majority of the work on attack detection leverages the established literature of fault detection [1],[2],[10],[11]. A fault detection approach uses an *estimator* to forecast the evolution of the system dynamics. When the residual (the difference between what is measured and the estimation), or some function of the residual, is larger than a predetermined threshold, an alarm is raised. Arguably the most insidious attacks are those that occur without our knowledge. Fault detectors impose limits on the attacker, if the attacker aims to avoid being identified. Beyond retooling these existing methods for the new attack detection context, a fundamental question is: given a chosen fault detection approach, how does this method constrain the influence of an attacker? More specifically, what is an attacker able to accomplish when a system employs certain fault detection procedure?

Different methodologies exist for evaluating the impact of attacks. Most of the existing work uses some measure

of state (or state estimate) deviation. In [2], the authors identify that if the attacker can take advantage of the zero dynamics of a (noise-free) input-output system, it can modify the system dynamics without reflecting its influence in the residual variables. This type of attacks are stealthy to any fault detector. A number of groups have studied the system response when the attacks are constrained by the detector. An important distinction between the collection of existing work – and the work discussed here – is the definition of how the attacker is constrained. We suggest the following terminology. While the term *stealthy attack* is used very broadly, we suggest that this refer to the zero-dynamics case, as discussed in [2], because these attacks do not propagate to the residual. Some work has investigated the case of system response due to what we here call *zero-alarm attacks*, i.e., attacks such that the detector threshold is never crossed [12]-[16]. Because real systems (with noise) always have a nonzero rate of false alarms raised by the detector, this attack model yields a relatively obvious attack signature because the alarms stop as soon as the attack starts. Other papers identify attacks that mimic the false alarm rate, thus making the alarm rate during the attack very close to the false alarm rate before the attack started [17],[18]. These attacks we call *hidden attacks* because although they do change the distribution of the residual, these changes are hidden from the way the detector evaluates the distribution. A majority of this work uses state bounds or steady-state limits to quantify the impact that an attacker can have. The exceptions to this are [17],[18], which quantify the reachable set of states and estimation errors when driven by the attack input.

This paper fuses several of these successful lines of research with a more strict interpretation of hidden attacks. The papers [17],[18] consider hidden attacks, however, they permit the alarm rate to change by a small value; the attacker capabilities that are derived are associated with this small deviation rather than the full scope of allowable attacks. Here, we fix the alarm rate exactly to study true hidden attacks (i.e., alarm rate exactly equal to the false alarm rate), and characterize the reachable sets on the estimation error dynamics associated with this broader definition of possible attack vectors. In this work, we characterize the *hidden reachable sets* that the attacker can induce through manipulation of sensor data. Because in general, it is quite difficult to compute these sets exactly, for given system dynamics and attack detection scheme, we derive *ellipsoidal bounds* on the hidden reachable sets using Linear Matrix Inequalities (LMIs) [19]. Then, we provide synthesis tools for minimizing these bounds (minimizing thus the hidden reachable set) by properly redesigning the detectors.

This work was supported by the National Research Foundation (NRF), Prime Minister’s Office, Singapore, under its National Cybersecurity R&D Programme (Award No. NRF2014NCR-NCR001-40) and administered by the National Cybersecurity R&D Directorate.

C. Murguia is with the Engineering Systems and Design Pillar, Singapore University of Technology and Design. J. Ruths is with the Departments of Mechanical and Systems Engineering, University of Texas at Dallas. emails: murguia\_rendon@sutd.edu.sg & jruths@utdallas.edu.

This builds off of our previous work in [18]. The strict interpretation of hidden attacks requires more direct handling of the effect of noise. To derive finite ellipsoidal bounds, we introduce the notion of  $p$ -probable reachable sets, which provides a nested set of ellipsoidal bounds based on the probability of the driving random sequences taking certain values. Because the derivation of the reachable set of states from the reachable set of estimation errors is captured in [18] (for a class of observer-based output feedback controllers), and similar techniques can be used in this paper, we report here only on estimation error reachable sets. Note that the problem formulation in this paper, while seemingly similar, requires an entirely different characterization from [18].

## II. SYSTEM DESCRIPTION & ATTACK DETECTION

We study LTI stochastic systems of the form:

$$\begin{cases} x(t_{k+1}) = Fx(t_k) + Gu(t_k) + v(t_k), \\ y(t_k) = Cx(t_k) + \eta(t_k), \end{cases} \quad (1)$$

with sampling time-instants  $t_k, k \in \mathbb{N}$ , state  $x \in \mathbb{R}^n$ , measured output  $y \in \mathbb{R}^m$ , control input  $u \in \mathbb{R}^l$ , matrices  $F$ ,  $G$ , and  $C$  of appropriate dimensions, and i.i.d. multivariate zero-mean Gaussian noises  $v \in \mathbb{R}^n$  and  $\eta \in \mathbb{R}^m$  with covariance matrices  $R_1 \in \mathbb{R}^{n \times n}$ ,  $R_1 \geq 0$  and  $R_2 \in \mathbb{R}^{m \times m}$ ,  $R_2 \geq 0$ , respectively. The initial state  $x(t_1)$  is assumed to be a Gaussian random vector with covariance matrix  $R_0 \in \mathbb{R}^{n \times n}$ ,  $R_0 \geq 0$ . The processes  $v(t_k)$ ,  $k \in \mathbb{N}$  and  $\eta(t_k)$ ,  $k \in \mathbb{N}$  and the initial condition  $x(t_1)$  are mutually independent. It is assumed that  $(F, G)$  is stabilizable and  $(F, C)$  is detectable. At the time-instants  $t_k, k \in \mathbb{N}$ , the output of the process  $y(t_k)$  is sampled and transmitted over a communication network. The received output  $\bar{y}(t_k)$  is used to compute control actions  $u(t_k)$  which are sent back to the process, see Fig. 1. The complete control-loop is assumed to be performed instantaneously, i.e., the sampling, transmission, and arrival time-instants are supposed to be equal. In this paper, we focus on attacks on sensor measurements. That is, in between transmission and reception of sensor data, an attacker may replace the signals coming from the sensors to the controller, see Fig. 1. After each transmission and reception, the attacked output  $\bar{y}$  takes the form:

$$\bar{y}(t_k) := y(t_k) + \delta(t_k) = Cx(t_k) + \eta(t_k) + \delta(t_k), \quad (2)$$

where  $\delta(t_k) \in \mathbb{R}^m$  denotes *additive sensor attacks*. Denote  $x_k := x(t_k)$ ,  $u_k := u(t_k)$ ,  $v_k := v(t_k)$ ,  $\bar{y}_k := \bar{y}(t_k)$ ,  $\eta_k := \eta(t_k)$ , and  $\delta_k := \delta(t_k)$ . Using this new notation, the attacked system is written as follows

$$\begin{cases} x_{k+1} = Fx_k + Gu_k + v_k, \\ \bar{y}_k = Cx_k + \eta_k + \delta_k. \end{cases} \quad (3)$$

### A. Observer

In order to estimate the state of the process, we use the following Luenberger observer [20]

$$\hat{x}_{k+1} = F\hat{x}_k + Gu_k + L(\bar{y}_k - C\hat{x}_k), \quad (4)$$

with estimated state  $\hat{x}_k \in \mathbb{R}^n$ ,  $\hat{x}_1 = E[x(t_1)]$ , where  $E[\cdot]$  denotes expectation, and observer gain matrix  $L \in \mathbb{R}^{n \times m}$ . Define the estimation error  $e_k := x_k - \hat{x}_k$ . Given the system

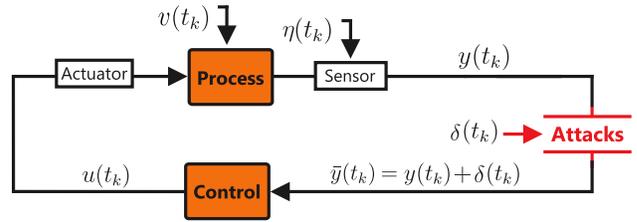


Fig. 1. Cyber-physical system under attacks on the sensor measurements.

dynamics (3) and the observer (4), the estimation error is governed by the following difference equation

$$e_{k+1} = (F - LC)e_k - L\eta_k - L\delta_k + v_k. \quad (5)$$

The pair  $(F, C)$  is detectable; hence, the observer gain  $L$  can be selected such that  $(F - LC)$  is Schur. Moreover, under detectability of  $(F, C)$ , if there are no attacks (i.e.,  $\delta_k = \mathbf{0}$ ), where  $\mathbf{0}$  denotes the zero matrix of appropriate dimensions, the covariance matrix  $P_k := E[e_k e_k^T]$  converges to steady state in the sense that  $\lim_{k \rightarrow \infty} P_k = P$  exists, see [21]. For a given  $L$  and  $\delta_k = \mathbf{0}$ , it can be verified that the asymptotic covariance matrix  $P = \lim_{k \rightarrow \infty} P_k$  is given by the solution  $P$  of the following Lyapunov equation:

$$(F - LC)P(F - LC)^T - P + R_1 + LR_2L^T = \mathbf{0}. \quad (6)$$

It is assumed that the system has reached steady state before an attack occurs.

### B. Residuals and Hypothesis Testing

In this manuscript, we characterize the effect that output injection attacks can induce in the system with being detected by *fault detection techniques*. The main idea behind fault detection theory is the use of an estimator to forecast the evolution of the system. If the difference between what it is measured and the estimation is larger than expected, there may be a fault in or attack on the system. Although the notion of residuals and model-based detectors is now routine in the fault detection literature, the primary focus has been on detecting and isolating failures that have known signatures in the degradation of measurement quality, i.e., faults with specific structures. Now, in the context of an intelligent adversarial attacker for which there is no known attack signature, new challenges arise to understand the effect that an adaptive intruder can have on the system without being detected. In this paper, we use the linear observer introduced in the previous section as our estimator. Define the *residual sequence*  $r_k, k \in \mathbb{N}$ , as

$$r_k := \bar{y}_k - C\hat{x}_k = Ce_k + \eta_k + \delta_k, \quad (7)$$

which evolves according to the difference equation:

$$\begin{cases} e_{k+1} = (F - LC)e_k - L\eta_k - L\delta_k + v_k, \\ r_k = Ce_k + \eta_k + \delta_k. \end{cases} \quad (8)$$

If there are no attacks, the steady state mean of  $r_k$  is

$$E[r_{k+1}] = CE[e_{k+1}] + E[\eta_{k+1}] = \mathbf{0}_{m \times 1}, \quad (9)$$

and its asymptotic covariance matrix is given by

$$\Sigma := E[r_{k+1}r_{k+1}^T] = CPC^T + R_2. \quad (10)$$

It is assumed that  $\Sigma \in \mathbb{R}^{m \times m}$  is positive definite. For this residual, we identify two hypotheses to be tested:  $\mathcal{H}_0$  the *normal mode* (no attacks) and  $\mathcal{H}_1$  the *faulty mode* (with faults/attacks). Then, we have

$$\mathcal{H}_0 : \begin{cases} E[r_k] = \mathbf{0}_{m \times 1}, \\ E[r_k r_k^T] = \Sigma, \end{cases} \quad \mathcal{H}_1 : \begin{cases} E[r_k] \neq \mathbf{0}_{m \times 1}, \text{ or} \\ E[r_k r_k^T] \neq \Sigma, \end{cases}$$

where  $\mathbf{0}_{m \times 1}$  denotes an  $m$ -dimensional vector composed of zeros only. In this manuscript, we use the chi-squared procedure for examining the residual and subsequently detecting attacks.

### C. Distance Measure and Chi-squared Procedure

The input to any detection procedure is a *distance measure*  $z_k \in \mathbb{R}$ ,  $k \in \mathbb{N}$ , i.e., a measure of how deviated the estimator is from the sensor measurements. We employ distance measures any time we test to distinguish between  $\mathcal{H}_0$  and  $\mathcal{H}_1$ . The chi-squared test uses a quadratic form on the residual as distance measure to test for substantial variations in mean and variance of the error between the measured output and the estimate. Consider the residual sequence  $r_k$ , (8), and its covariance matrix  $\Sigma$ , (10). The chi-squared procedure is defined as follows.

---

#### Chi-squared procedure:

$$\text{If } z_k := r_k^T \Sigma^{-1} r_k > \alpha, \quad \tilde{k} = k. \quad (11)$$

**Design parameter:** threshold  $\alpha \in \mathbb{R}_{>0}$ .

**Output:** alarm time(s)  $\tilde{k}$ .

---

Thus, the procedure is designed so that alarms are triggered if  $z_k$  exceeds the threshold  $\alpha$ . The normalization by  $\Sigma^{-1}$  makes setting the value of the threshold  $\alpha$  system independent. This quadratic expression leads to a sum of the squares of  $m$  normally distributed random variables which implies that the distance measure  $z_k$  follows a chi-squared distribution with  $m$  degrees of freedom, see, e.g., [22] for details.

### D. False Alarms

The occurrence of an alarm in the chi-squared procedure when there are no attacks to the CPS is referred to as a false alarm. The threshold  $\alpha$  must be selected to fulfill a *desired false alarm rate*  $\mathcal{A}^*$ . Let  $\mathcal{A} \in [0, 1]$  denote the *false alarm rate* of the chi-squared procedure defined as the expected proportion of observations which are false alarms, i.e.,  $\mathcal{A} := \text{pr}[z_k \geq \alpha]$ , where  $\text{pr}[\cdot]$  denotes probability, see [23] and [24].

**Proposition 1** [13]. *Assume that there are no attacks on the system and consider the chi-squared procedure (11) with residual  $r_k \sim \mathcal{N}(\mathbf{0}, \Sigma)$  and threshold  $\alpha \in \mathbb{R}_{>0}$ . Let  $\alpha = \alpha^* := 2\text{P}^{-1}(\frac{m}{2}, 1 - \mathcal{A}^*)$ , where  $\text{P}^{-1}(\cdot, \cdot)$  denotes the inverse regularized lower incomplete gamma function (see [22]), then  $\mathcal{A} = \mathcal{A}^*$ .*

## III. HIDDEN REACHABLE SETS

In this section, we provide tools for *quantifying* (for given  $L$ ) and *minimizing* (by selecting  $L$ ) the impact of the

attack sequence  $\delta_k$  on the estimation error  $e_k$  when the chi-squared procedure is used for attack detection. To quantify the effect of attacks, we need to introduce some measure of impact. However, because malicious adversaries may launch any arbitrary attack, we need a measure which can capture all possible trajectories that the attacker can induce in the estimation error dynamics, given how it accesses the dynamics (i.e., through residual variables by tampering with sensor measurements). We propose to use the *reachable set* of the attack sequence  $\delta_k$  as our measure of impact. We are interested in *attacks that do not change the false alarm rate* of the detector  $\mathcal{A}$ , i.e.,  $\bar{\mathcal{A}} = \mathcal{A}$ , where  $\bar{\mathcal{A}}$  denotes the alarm rate under the attacker's action. This class of attacks is what we refer to as *hidden attacks* and the trajectories that hidden attacks can induce in the system are referred to as *hidden reachable sets*. In this section, we provide tools based on Linear Matrix Inequalities (LMIs) for computing outer ellipsoidal bounds on the hidden reachable sets induced by the attack sequence  $\delta_k$  given the system dynamics, the chi-squared procedure, the noise, and the false alarm rate  $\mathcal{A}$ .

### A. Attack Model and Hidden Reachable Sets

We assume that the attacker has perfect knowledge of the system dynamics, the observer, measurements, and detection procedure (chi-squared). It is further assumed that all the sensors can be compromised by the attacker at each time step (the case where not all the sensors are attacked is left as future work). By considering this strong, worst-case attacker, we are able to construct an upper bound on the abilities of the attacker. Consider the estimation error dynamics (8), the residual sequence  $r_k = Ce_k + \eta_k + \delta_k$ , and the distance measure

$$z_k = \|\Sigma^{-\frac{1}{2}} r_k\|^2 = \|\Sigma^{-\frac{1}{2}} (Ce_k + \eta_k + \delta_k)\|^2, \quad (12)$$

where  $\Sigma^{-\frac{1}{2}}$  denotes the symmetric squared root matrix of  $\Sigma^{-1}$ . The set of feasible attack sequences that the opponent can launch while satisfying  $\bar{\mathcal{A}} = \mathcal{A}$  can be written as the following constrained control problem on  $\delta_k$ :

$$\left\{ \delta_k \in \mathbb{R}^m \mid \begin{array}{l} e_k \text{ satisfies (8), and} \\ \text{pr}[\|\Sigma^{-\frac{1}{2}} (Ce_k + \eta_k + \delta_k)\|^2 > \alpha] = \mathcal{A}, \end{array} \right\}, \quad (13)$$

for  $k \in \mathbb{N}$ . We are interested in the error trajectories that the attacker can induce in the system restricted to satisfy (13). Note that, as long as  $\bar{\mathcal{A}} = \mathcal{A}$ , the attacker may induce any arbitrary random sequence  $\delta_k$ . This and the fact that  $v_k$  and  $\eta_k$  are Gaussian (thus having infinite support) imply that deterministic reachable sets induced by  $\delta_k$  and the noise sequences are generally unbounded. To overcome this obstacle, we introduce the notion of *p-probable hidden reachable sets*  $\mathcal{R}_\alpha^p$ . Define  $\zeta_k := \Sigma^{-\frac{1}{2}} (Ce_k + \eta_k + \delta_k)$  and note that the estimation error dynamics (8) can be written in terms of  $\zeta_k$  as:

$$\begin{cases} e_{k+1} = Fe_k - L\Sigma^{\frac{1}{2}}\zeta_k + v_k, \\ \zeta_k = \Sigma^{-\frac{1}{2}}(Ce_k + \eta_k + \delta_k). \end{cases} \quad (14)$$

For given false alarm rate  $\mathcal{A}$  and probability  $p \in (0, 1)$ , the *p-probable hidden reachable set* of the attack sequence  $\delta_k$  in (14),  $\mathcal{R}_\alpha^p$ , is defined as the set of  $e_k \in \mathbb{R}^n$ ,  $k \in \mathbb{N}$  that can

be reached from the origin  $e_1 = \mathbf{0}$  due to the the attacker's action  $\delta_k$  restricted to satisfy  $\bar{\mathcal{A}} = \mathcal{A}$  and

$$\text{pr}[\|\zeta_k\|^2 \leq \bar{\zeta}_p] = \text{pr}[\|v_k\|^2 \leq \bar{v}_p] = p, \quad (15)$$

for some constants  $\bar{\zeta}_p, \bar{v}_p \in \mathbb{R}_{>0}$ , i.e.,

$$\mathcal{R}_\alpha^p := \left\{ e_k \in \mathbb{R}^n \left| \begin{array}{l} e_1 = \mathbf{0}, \\ e_k, \delta_k, v_k \text{ satisfy (13)-(15)}, \end{array} \right. \right\}. \quad (16)$$

By restricting the probabilities in (15), we are delimiting the support of the attack and noise sequences to compact sets. Then, the  $p$ -probable hidden reachable sets correspond to the trajectories of the system when the driving random sequences are restricted to satisfy  $\bar{\mathcal{A}} = \mathcal{A}$  and (15). For delimited  $v_k$  and  $\delta_k$ , we can characterize reachable sets using deterministic tools. In general, it is analytically intractable to compute  $\mathcal{R}_\alpha^p$  exactly. Instead, using LMIs, for some positive definite matrix  $\mathcal{P}_\alpha^p \in \mathbb{R}^{n \times n}$ , we derive outer ellipsoidal bounds of the form  $\mathcal{E}_\alpha^p := \{e_k \in \mathbb{R}^n | e_k^T \mathcal{P}_\alpha^p e_k \leq 1\}$  containing  $\mathcal{R}_\alpha^p$ .

**Remark 1** Note, from (14), that if for some  $k = k^*$ ,  $e_{k^*} \neq \mathbf{0}$  and  $\rho[F] > 1$ , where  $\rho[\cdot]$  denotes spectral radius, then  $\|e_k\|$  diverges to infinity as  $k \rightarrow \infty$  for any non-stabilizing  $\zeta_k$ . That is,  $\mathcal{R}_\alpha^p$  is unbounded if the system is open-loop unstable. If  $\rho[F] \leq 1$ , then  $\|e_k\|$  may or may not diverge to infinity depending on algebraic and geometric multiplicities of the eigenvalues with unit modulus of  $F$  (a known fact from stability of LTI systems), see [21] for details.

Given Remark 1, in what follows, we consider open-loop stable systems ( $\rho[F] < 1$ ). The following result is used to compute the ellipsoidal bounds  $\mathcal{E}_\alpha^p$ .

**Lemma 1** [25] Let  $\xi_k \in \mathbb{R}^n$ ,  $\xi_1 = \mathbf{0}$ ,  $V_k := \xi_k^T \mathcal{P} \xi_k$ , for some positive definite matrix  $\mathcal{P} \in \mathbb{R}^{n \times n}$ , and  $\omega_k^T \omega_k \leq \bar{\omega}$ ,  $\bar{\omega} \in \mathbb{R}_{>0}$ . If there exists a constant  $b \in (0, 1)$  such that

$$V_{k+1} - bV_k - \frac{1-b}{\bar{\omega}} \omega_k^T \omega_k \leq 0, \forall k \in \mathbb{N}, \quad (17)$$

then,  $V_k = \xi_k^T \mathcal{P} \xi_k \leq 1$ .

*B. Case 1:  $p \in [0, 1 - \mathcal{A}]$*

Because the attack sequence is restricted to satisfy (13), we start computing the ellipsoidal bounds corresponding to  $p = 1 - \mathcal{A}$ , i.e.,  $\mathcal{E}_\alpha^{1-\mathcal{A}}$ . It is easy to verify using Lemma 1 that  $\mathcal{E}_\alpha^p \subseteq \mathcal{E}_\alpha^{1-\mathcal{A}}$  for  $p \in [0, 1 - \mathcal{A}]$  because  $\bar{\zeta}_p \leq \bar{\zeta}_{1-\mathcal{A}} = \alpha$  and  $\bar{v}_p \leq \bar{v}_{1-\mathcal{A}}$  in (15); i.e., all  $p$ -probable ellipsoidal bounds for  $p \in [0, 1 - \mathcal{A}]$  lie within the  $1 - \mathcal{A}$ -probable ellipsoidal bound. It follows that  $\mathcal{R}_\alpha^p \subseteq \mathcal{E}_\alpha^{1-\mathcal{A}}$  for  $p \in [0, 1 - \mathcal{A}]$ , i.e., for  $p$  in this interval, we only need to compute the ellipsoidal bound corresponding to  $p = 1 - \mathcal{A}$ . Characterizing  $p$ -probable sets for small  $p$  values is of little interest because they do not provide a informative bound on system trajectories (since the smaller  $p$  is, the more trajectories lie outside the  $p$ -probable ellipsoidal bound). We work with the data available in this setting, namely the number of alarms raised by the detector, to bound the most informative  $p = 1 - \mathcal{A}$  probable reachable set; in Case 2, we extend these results for larger  $p$  values.

**Theorem 1** For given system matrix  $F$ , observer gain  $L$ , residual covariance matrix  $\Sigma$ , and false alarm rate  $\mathcal{A}$ , consider the set  $\mathcal{R}_\alpha^{1-\mathcal{A}}$  in (16). If there exists a positive

definite matrix  $\mathcal{P} \in \mathbb{R}^{n \times n}$  and  $b \in (0, 1)$  satisfying the following matrix inequality:

$$\begin{bmatrix} b\mathcal{P} & F^T \mathcal{P} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathcal{P} F & \mathcal{P} & \mathcal{P} & -\mathcal{P} L \Sigma^{\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & -\Sigma^{\frac{1}{2}} L^T \mathcal{P} & \frac{1-b}{\bar{\omega}} I & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \frac{1-b}{\bar{\omega}} I & I \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & I \end{bmatrix} \geq \mathbf{0}; \quad (18)$$

for  $\bar{\omega} = \alpha + \bar{v}_{1-\mathcal{A}}$ ; then,  $\mathcal{R}_\alpha^{1-\mathcal{A}} \subseteq \mathcal{E}_\alpha^{1-\mathcal{A}}$  with  $\mathcal{P}_\alpha^{1-\mathcal{A}} = \mathcal{P}$ , i.e., the  $(1 - \mathcal{A})$ -probable hidden reachable set is contained in the ellipsoid  $\mathcal{E}_\alpha^{1-\mathcal{A}} = \{e_k \in \mathbb{R}^n | e_k^T \mathcal{P}_\alpha^{1-\mathcal{A}} e_k \leq 1\}$ .

**Proof:** For a positive definite matrix  $\mathcal{P} \in \mathbb{R}^{n \times n}$ , consider the function  $V_k := e_k^T \mathcal{P} e_k$ , then, from (16), inequality (17) takes the form:

$$\begin{aligned} &= -\vartheta_k^T \begin{bmatrix} b\mathcal{P} - F^T \mathcal{P} F & F^T \mathcal{P} L \Sigma^{\frac{1}{2}} & -F^T \mathcal{P} \\ \Sigma^{\frac{1}{2}} L^T \mathcal{P} F & \frac{1-b}{\bar{\omega}} I - \Sigma^{\frac{1}{2}} L^T \mathcal{P} L \Sigma^{\frac{1}{2}} & \Sigma^{\frac{1}{2}} L^T \mathcal{P} \\ -\mathcal{P} F & \mathcal{P} L \Sigma^{\frac{1}{2}} & \frac{1-b}{\bar{\omega}} I - \mathcal{P} \end{bmatrix} \vartheta_k \\ &=: -\vartheta_k^T \mathcal{Q}_e \vartheta_k \leq 0, \end{aligned}$$

where  $\vartheta := (e_k^T, \zeta_k^T, v_k^T)^T$ . The above inequality is satisfied if and only if  $\mathcal{Q}_e \geq \mathbf{0}$ . Matrix  $\mathcal{Q}_e$  can be written as the Schur complement of a higher dimensional matrix  $\mathcal{Q}'_e$ ; hence,  $\mathcal{Q}_e \geq \mathbf{0} \leftrightarrow \mathcal{Q}'_e \geq \mathbf{0}$ , i.e.,

$$\mathcal{Q}_e \geq \mathbf{0} \leftrightarrow \begin{bmatrix} b\mathcal{P} & \mathbf{0} & \mathbf{0} & F^T \mathcal{P} & \mathbf{0} \\ \mathbf{0} & \frac{1-b}{\bar{\omega}} I & \mathbf{0} & -\Sigma^{\frac{1}{2}} L^T \mathcal{P} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{1-b}{\bar{\omega}} I & \mathcal{P} & \mathbf{0} \\ \mathcal{P} F - \mathcal{P} L \Sigma^{\frac{1}{2}} & \mathcal{P} & \mathcal{P} & \mathcal{P} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & I \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \geq \mathbf{0}. \quad (19)$$

Finally, inequality (18) follows from (19) by a simple re-ordering of rows and columns. The result follows now from Lemma 1 by taking  $\mathcal{P}_\alpha^{1-\mathcal{A}} = \mathcal{P}$  and  $\bar{\omega} = \alpha + \bar{v}_{1-\mathcal{A}}$ . ■

The result in Theorem 1 provides a tool for computing ellipsoidal bounds on  $\mathcal{R}_\alpha^{1-\mathcal{A}}$ . To make the bounds most useful, we next construct ellipsoids with minimal volume, i.e., the tightest possible ellipsoid bounding  $\mathcal{R}_\alpha^{1-\mathcal{A}}$ . In this case, we have to minimize  $\det \mathcal{P}^{-1}$  subject to (18) (because  $\det \mathcal{P}^{-1}$  is proportional to the volume of  $e_k^T \mathcal{P} e_k = 1$ ). This is formally stated in the following corollary of Theorem 1, see [19] for further details.

**Corollary 1** For given matrices  $(F, L, \Sigma)$ , false alarm rate  $\mathcal{A}$ , and  $b \in (0, 1)$ , the solution  $\mathcal{P}$  of the following convex optimization:

$$\begin{cases} \min_{\mathcal{P}} & -\log \det \mathcal{P}, \\ \text{s.t.} & \mathcal{P} > \mathbf{0} \text{ and (18)}, \end{cases} \quad (20)$$

for  $\bar{\omega} = \alpha + \bar{v}_{1-\mathcal{A}}$ , minimizes the volume of the ellipsoid  $\mathcal{E}_\alpha^{1-\mathcal{A}}$  (with  $\mathcal{P}_\alpha^{1-\mathcal{A}} = \mathcal{P}$ ) bounding  $\mathcal{R}_\alpha^{1-\mathcal{A}}$ .

See [26] for an example of how to solve (26) using YALMIP.

As we now move toward redesigning  $L$  to minimize the ellipsoids, we note that as  $\|L\| \rightarrow 0$ , the volume of  $\mathcal{E}_\alpha^{1-\mathcal{A}}$  goes to zero because the attack-dependent term in (14),  $L \Sigma^{\frac{1}{2}} \zeta_k$ , vanishes. In other words, without any other

considered criteria, the observer gain leading to the minimum volume ellipsoid is trivially given by  $L = \mathbf{0}$ . While this is effective at eliminating the impact of the attacker, it implies that we discard the observer altogether and, therefore, forfeit any ability to build a reliable estimate of the system state. If we impose a performance criteria that the observer must satisfy in the attack-free case (e.g., convergence speed, noise-output gain, and minimum asymptotic variance), it has to be added into the minimization problem (26) so as to minimize the volume of  $\mathcal{E}_\alpha^{1-\mathcal{A}}$  while still achieving the observer performance in the attack-free case. For completeness, in the following proposition, we provide an LMI criteria for ensuring that the  $H_\infty$  gain from the noise to the residual  $r_k$  in (8) is less than or equal to some  $\gamma \in \mathbb{R}_{>0}$ . Then, using this criteria and Theorem 1, we provide a synthesis tool for minimizing the volume of  $\mathcal{E}_\alpha^{1-\mathcal{A}}$  while ensuring a desired  $H_\infty$  performance in the attack-free case.

**Proposition 2** For given matrices  $(F, C, L)$ , if there exist a positive definite matrix  $\mathcal{P} \in \mathbb{R}^{n \times n}$  and constant  $\gamma \in \mathbb{R}_{>0}$  satisfying the following matrix inequality:

$$\begin{bmatrix} \mathcal{P} & \mathbf{0} & \mathbf{0} & (F-LC)^T \mathcal{P} & C^T \\ \mathbf{0} & \gamma^2 I & \mathbf{0} & -L^T \mathcal{P} & I \\ \mathbf{0} & \mathbf{0} & \gamma^2 I & \mathcal{P} & \mathbf{0} \\ \mathcal{P}(F-LC) & -\mathcal{P}L & \mathcal{P} & \mathcal{P} & \mathbf{0} \\ C & I & \mathbf{0} & \mathbf{0} & I \end{bmatrix} \geq \mathbf{0}, \quad (21)$$

then, the  $H_\infty$  gain from the noise  $\nu_k := (\eta_k^T, v_k^T)^T$  to the residual  $r_k = Ce_k + \eta_k$  of the estimation error dynamics (8) is less than or equal to  $\gamma$ .

The proof of Proposition 2 is omitted here due to the page limit. However, this is a standard result and details about the proof can be found in, e.g., [9] and references therein. In the following corollary of Theorem 1 and Proposition 2, we formulate the optimization problem for designing the observer gain  $L$  such that the volume of the ellipsoid  $\mathcal{E}_\alpha^{1-\mathcal{A}}$  is minimized and a desired  $H_\infty$  performance is achieved in the attack-free case.

**Corollary 2** For given system matrices  $(F, C)$ , residual covariance matrix  $\Sigma$ , false alarm rate  $\mathcal{A}$ ,  $b \in (0, 1)$ , and  $\gamma \in \mathbb{R}_{>0}$ , if there exist matrices  $\mathcal{P} \in \mathbb{R}^{n \times n}$  and  $M \in \mathbb{R}^{n \times m}$  solution to the following convex optimization:

$$\begin{cases} \min_{\mathcal{P}, M} & -\log \det \mathcal{P}, \\ \text{s.t. } \mathcal{P} > \mathbf{0}, & \begin{bmatrix} b\mathcal{P} & F^T \mathcal{P} & \mathbf{0} & \mathbf{0} & \mathbf{0} \mathbf{0} \\ \mathcal{P}F & \mathcal{P} & \mathcal{P} & -M\Sigma^{\frac{1}{2}} & \mathbf{0} \mathbf{0} \\ \mathbf{0} & \mathcal{P} & \frac{1-b}{\bar{\omega}} I & \mathbf{0} & \mathbf{0} \mathbf{0} \\ \mathbf{0} & -\Sigma^{\frac{1}{2}} M^T & \mathbf{0} & \frac{1-b}{\bar{\omega}} I & \mathbf{0} \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & I \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & I \end{bmatrix} \geq \mathbf{0}, \text{ and} \\ & \begin{bmatrix} \mathcal{P} & \mathbf{0} & \mathbf{0} & F^T \mathcal{P} - C^T M^T & C^T \\ \mathbf{0} & \gamma^2 I & \mathbf{0} & -M^T & I \\ \mathbf{0} & \mathbf{0} & \gamma^2 I & \mathcal{P} & \mathbf{0} \\ \mathcal{P}F - MC - M & \mathcal{P} & \mathcal{P} & \mathcal{P} & \mathbf{0} \\ C & I & \mathbf{0} & \mathbf{0} & I \end{bmatrix} \geq \mathbf{0}, \end{cases} \quad (22)$$

for  $\bar{\omega} = \alpha + \bar{v}_{1-\mathcal{A}}$ ; then, the observer gain  $L = \mathcal{P}^{-1}M$  minimizes the volume of the ellipsoid  $\mathcal{E}_\alpha^{1-\mathcal{A}}$  (with  $\mathcal{P}_\alpha^{1-\mathcal{A}} =$

$\mathcal{P}$ ) bounding  $\mathcal{R}_\alpha^{1-\mathcal{A}}$  and guarantees that the  $H_\infty$  gain from the noise  $\nu_k = (\eta_k^T, v_k^T)^T$  to the residual  $r_k$  of (8) is less than or equal to  $\gamma$  in the attack-free case.

**Proof:** This follows from Theorem 1, Proposition 2, and the linearizing change of variables  $M = \mathcal{P}L$ .  $\blacksquare$

C. Case 2:  $p \in (1 - \mathcal{A}, 1]$

Note that, for  $p \in (1 - \mathcal{A}, 1]$ ,  $\bar{\zeta}_p > \bar{\zeta}_{1-\mathcal{A}} = \alpha$  according to (15). Then, we can write  $\bar{\zeta}_p = \alpha + \epsilon_p$  and  $\text{pr}[\|\zeta_k\|^2 \leq \alpha + \epsilon_p] = 1 - \mathcal{A} + a_p$ , for some  $\epsilon_p \in (0, \infty)$  and  $a_p \in (0, \mathcal{A}]$ . To be able to compute ellipsoidal bounds, the constant  $\epsilon_p$  corresponding to a given probability  $1 - \mathcal{A} + a_p$  is required. If  $\epsilon_p$  is available, we can restrict  $\zeta_k$  to compact sets as in Case 1. Note, however, that the distribution of the attack sequence  $\delta_k$  (and thus the one of  $\zeta_k$ ) is generally unknown. Actually, the attacker may induce any arbitrary (and possibly) non-stationary random sequence  $\zeta_k$  in (14) as long as  $\bar{\mathcal{A}} = \mathcal{A}$ . Nevertheless, we can obtain bounds on  $\epsilon_p$  using Markov's inequality [22] to link the statistical properties of  $\zeta_k$  with  $\epsilon_p$ . This is stated in the following proposition.

**Proposition 3** Denote  $\mathcal{M}_k := E[\zeta_k \zeta_k^T]$  and  $\mu_k := E[\zeta_k]$ . For given false alarm rate  $\mathcal{A}$ , probability  $p = 1 - \mathcal{A} + a_p$ , and  $a_p \in (0, \mathcal{A})$ , the following is satisfied:

$$\left\{ \begin{array}{l} \text{pr}[\|\zeta_k\|^2 \leq \alpha + \epsilon_p] \in [1 - \mathcal{A} + a_p, 1], \\ \text{for all } \epsilon_p \geq \underline{\epsilon}_p := \frac{\text{tr}[\mathcal{M}_k] + \mu_k^T \mu_k}{\mathcal{A} - a_p} - \alpha. \end{array} \right. \quad (23a)$$

$$\left\{ \begin{array}{l} \text{pr}[\|\zeta_k\|^2 \leq \alpha + \epsilon_p] \in [1 - \mathcal{A} + a_p, 1], \\ \text{for all } \epsilon_p \geq \underline{\epsilon}_p := \frac{\text{tr}[\mathcal{M}_k] + \mu_k^T \mu_k}{\mathcal{A} - a_p} - \alpha. \end{array} \right. \quad (23b)$$

**Proof:** The probability  $\text{pr}[\|\zeta_k\|^2 \leq \alpha + \epsilon_p]$  can be written as  $\text{pr}[\|\zeta_k\|^2 \leq \alpha + \epsilon_p] = 1 - \text{pr}[\|\zeta_k\|^2 > \alpha + \epsilon_p]$ . Then, using Markov's inequality [22], we can write the following

$$\text{pr}[\|\zeta_k\|^2 > \alpha + \epsilon_p] = 1 - \text{pr}[\|\zeta_k\|^2 \leq \alpha + \epsilon_p] \leq \frac{E[\|\zeta_k\|^2]}{\alpha + \epsilon_p}.$$

Therefore, if  $\epsilon_p$  satisfies  $E[\|\zeta_k\|^2]/(\alpha + \epsilon_p) \leq \mathcal{A} - a_p$ , then  $\text{pr}[\|\zeta_k\|^2 > \alpha + \epsilon_p] \leq \mathcal{A} - a_p$  and hence  $\text{pr}[\|\zeta_k\|^2 \leq \alpha + \epsilon_p] \in [1 - \mathcal{A} + a_p, 1]$ . The expectation of the quadratic form  $\zeta_k^T \zeta_k$  is given by  $E[\|\zeta_k\|^2] = E[\zeta_k^T \zeta_k] = \text{tr}[\mathcal{M}_k] + \mu_k^T \mu_k$  [22]; then,  $E[\|\zeta_k\|^2]/(\alpha + \epsilon_p) \leq \mathcal{A} - a_p$  is satisfied for all  $\epsilon_p \geq \underline{\epsilon}_p$  with  $\underline{\epsilon}_p$  as defined in (23b), and the assertion follows.  $\blacksquare$

Using Proposition 3, for given false alarm rate  $\mathcal{A}$  and probability  $p = 1 - \mathcal{A} + a_p \in (1 - \mathcal{A}, 1]$ ,  $a_p \in (0, \mathcal{A})$ , we can characterize  $p$ -probable hidden reachable sets,  $\tilde{\mathcal{R}}_\alpha^p$ , by using the lower bounds on  $\text{pr}[\|\zeta_k\|^2 \leq \alpha + \epsilon_p]$  and  $\epsilon_p$  in (23). Specifically, for  $p > 1 - \mathcal{A}$ , the set  $\tilde{\mathcal{R}}_\alpha^p$  of the sequence  $\delta_k$  is defined as the set of  $e_k \in \mathbb{R}^n$  that can be reached from  $e_1 = \mathbf{0}$  restricted to satisfy  $\bar{\mathcal{A}} = \mathcal{A}$  and

$$\left\{ \begin{array}{l} \text{pr}[\|v_k\|^2 \leq \bar{v}_p] = 1 - \mathcal{A} + a_p \text{ and} \\ \epsilon_p = \underline{\epsilon}_p \rightarrow \text{pr}[\|\zeta_k\|^2 \leq \alpha + \epsilon_p] \in [1 - \mathcal{A} + a_p, 1], \end{array} \right. \quad (24)$$

for some constant  $\bar{v}_p \in \mathbb{R}_{>0}$  and  $\underline{\epsilon}_p$  as defined in (24), i.e.,

$$\tilde{\mathcal{R}}_\alpha^p := \left\{ e_k \in \mathbb{R}^n \mid \begin{array}{l} e_1 = \mathbf{0}, \\ e_k, \zeta_k, v_k, \text{ satisfy (13)-(14),(24),} \end{array} \right\}. \quad (25)$$

**Remark 2** For a  $p$ -probable reachable set, we select  $a_p$  such that  $p = 1 - \mathcal{A} + a_p$ , then determine  $\epsilon_p$  using (23b). Note that, because the attacker can induce an attack sequence with

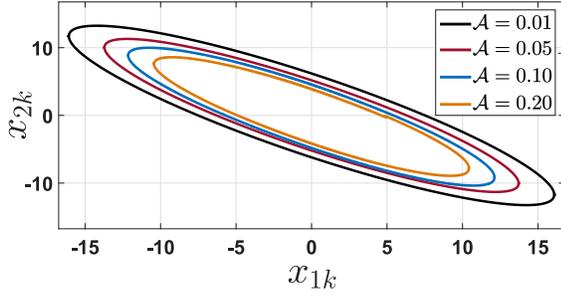


Fig. 2. Ellipsoid  $\mathcal{E}_\alpha^{1-\mathcal{A}}$  for different values of false alarm rate  $\mathcal{A}$ .

arbitrarily large covariance  $\mathcal{M}_k$  and mean  $\mu_k$ , the lower bound on  $\epsilon_p$ ,  $\underline{\epsilon}_p$ , in (23b) can be made arbitrarily large for any  $a_p$ . Therefore, if  $\delta_k$  (and thus  $\zeta_k$ ) is only restricted to satisfy  $\bar{A} = \mathcal{A}$ , the opponent can induce arbitrarily large reachable sets  $\tilde{\mathcal{R}}_\alpha^p$ .

Remark 2 implies that if we only monitor the alarms raised by the detector, the attacker can inject arbitrarily large signals in the residual sequence  $r_k$  without changing the alarm rate. Consequently, the sets  $\tilde{\mathcal{R}}_\alpha^p$  can be made arbitrarily large for arbitrarily small  $a_p$ . If we place additional assumptions on the attacker, namely that the mean and covariance of the attack sequence  $\zeta_k$  are finite, the reachable sets will be bounded by Proposition 3. In particular, if we assume the attacker maintains the mean and covariance of the attack-free scenario, i.e.,  $E[\zeta_k] = \mu_k = \mathbf{0}$  and  $E[\zeta_k \zeta_k^T] = \mathcal{M}_k = I_m$ , then  $\underline{\epsilon}_p = \frac{m}{\mathcal{A}-a_p} - \alpha$ . Hence, if in addition to imposing  $\bar{A} = \mathcal{A}$ , the attack is restricted to keep the statistical properties of  $\zeta_k$  in the attack-free case, i.e.,  $\mu_k = \mathbf{0}$  and  $\mathcal{M}_k = I_m$ , the reachable sets  $\tilde{\mathcal{R}}_\alpha^p$  are bounded for each  $a_p \in (0, \mathcal{A})$  (because  $\underline{\epsilon}_p$  is bounded); and therefore, in this case, we can compute ellipsoidal bounds on  $\tilde{\mathcal{R}}_\alpha^p$ . This additional assumption could be enforced by adding detectors that identify anomalies in the sample mean and sample covariance of the residual. Such detectors would force the attacker to avoid arbitrarily large attack values in order to avoid detection by these additional mean and covariance detectors.

As before, we characterize, for some positive definite matrix  $\tilde{\mathcal{P}}_\alpha^p \in \mathbb{R}^{n \times n}$ , outer ellipsoidal bounds of the form  $\tilde{\mathcal{E}}_\alpha^p := \{e_k \in \mathbb{R}^n | e_k^T \tilde{\mathcal{P}}_\alpha^p e_k \leq 1\}$  containing  $\tilde{\mathcal{R}}_\alpha^p$ . The results corresponding to Theorem 1, and Corollary 1 for Case 1 are stated in the following corollary.

**Corollary 3** For given false alarm rate  $\mathcal{A}$ , probability  $p = 1 - \mathcal{A} + a_p$ ,  $a_p \in (0, \mathcal{A})$ , threshold  $\epsilon_p = \underline{\epsilon}_p = \frac{m}{\mathcal{A}-a_p} - \alpha$ , and matrices  $(F, L, \Sigma)$ , consider the set  $\tilde{\mathcal{R}}_\alpha^p$  in (25). Then, for given  $b \in (0, 1)$ , if there exists a matrix  $\mathcal{P} \in \mathbb{R}^{n \times n}$  solution of the following convex optimization:

$$\begin{cases} \min_{\mathcal{P}} & -\log \det \mathcal{P}, \\ \text{s.t.} & \mathcal{P} > 0 \text{ and (18),} \end{cases} \quad (26)$$

for  $\bar{\omega} = \alpha + \underline{\epsilon}_p + \bar{v}_p$ ; then,  $\tilde{\mathcal{R}}_\alpha^p \subseteq \tilde{\mathcal{E}}_\alpha^p$  (with  $\tilde{\mathcal{P}}_\alpha^p = \mathcal{P}$ ) and  $\tilde{\mathcal{E}}_\alpha^p$  has minimum volume, i.e., the  $p$ -probable hidden reachable set  $\tilde{\mathcal{R}}_\alpha^p$  is contained in the minimum volume ellipsoid  $\mathcal{E}_\alpha^p = \{e_k \in \mathbb{R}^n | e_k^T \mathcal{P}_\alpha^p e_k \leq 1\}$ .

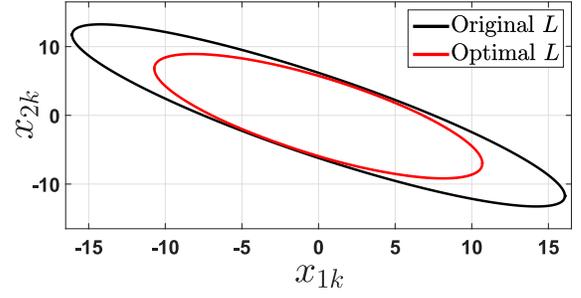


Fig. 3. The improvement in the  $(1 - \mathcal{A})$ -probable hidden reachable set ellipsoidal bound  $\mathcal{E}_\alpha^{1-\mathcal{A}}$ , for  $\mathcal{A} = 0.01$ , through application of Corollary 2 to design the optimal observer gain.

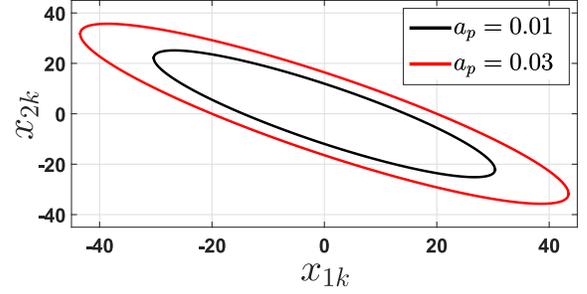


Fig. 4. Ellipsoidal bound  $\tilde{\mathcal{E}}_\alpha^p$  for different values of  $a_p$  obtained using Corollary 3.

A result for redesigning the observer gain for minimizing the volume of the above ellipsoids, as in Corollary 2 for Case 1, can be stated in a similar manner as the corollary above; however, this is omitted here due to the page limit.

#### IV. SIMULATION EXPERIMENTS

Consider the closed-loop system (3)-(4) with matrices:

$$\begin{cases} F = \begin{pmatrix} 0.84 & 0.23 \\ -0.47 & 0.12 \end{pmatrix}, G = \begin{pmatrix} 0.07 \\ 0.23 \end{pmatrix}, C = (1 \ 0), \\ L = \begin{pmatrix} 1.16 \\ -0.69 \end{pmatrix}, R_1 = \begin{pmatrix} 0.45 & -0.11 \\ -0.11 & 0.45 \end{pmatrix}, \\ R_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, R_2 = 1, \Sigma = 3.26. \end{cases} \quad (27)$$

We start with Case 1. Using Proposition 2, the observer gain  $L$  is designed such that the  $H_\infty$  gain from the noise to the residual  $r_k$  of (8) is less than or equal to  $\gamma = 1.86$  in the attack-free case. Consider the false alarm rates  $\mathcal{A} = \{0.01, 0.05, 0.10, 0.20\}$  and the corresponding  $\alpha = \{6.63, 3.84, 2.70, 1.64\}$ , obtained using Proposition 1. The thresholds  $\bar{v}_{1-\mathcal{A}}$  in (15) are computed such that  $\text{pr}\{\|v_k\|^2 \leq \bar{v}_{1-\mathcal{A}}\} = 1 - \mathcal{A}$ . Because the entries on the diagonal of  $R_1$  are equal and  $v_k \sim \mathcal{N}(\mathbf{0}, R_1)$ , the random sequence  $\|v_k\|^2$ ,  $k \in \mathbb{N}$  follows a gamma distribution,  $\Gamma(\kappa, \theta)$ , with shape parameter  $\kappa = 1$  and scale parameter  $\theta = 0.90$ , see [22]. It follows that, for these  $\mathcal{A}$ ,  $\bar{v}_{1-\mathcal{A}} = \{4.14, 2.69, 2.07, 1.44\}$ . For these values of  $\bar{v}_{1-\mathcal{A}}$  and  $\alpha$ , in Figure 2, we depict the ellipsoidal bounds  $\mathcal{E}_\alpha^{1-\mathcal{A}}$  on the  $(1 - \mathcal{A})$ -probable hidden reachable sets  $\tilde{\mathcal{R}}_\alpha^{1-\mathcal{A}}$  obtained using Theorem 1 and Corollary 1. Next, for  $\mathcal{A} = 0.01$ , using Corollary 2, we redesign the observer gain

$L$  to minimize the volume of  $\mathcal{E}_\alpha^{1-A}$  while maintaining the  $H_\infty$  performance below  $\gamma = 1.86$ . The obtained optimal ellipsoidal bound,  $\mathcal{E}_\alpha^{1-A}$ , is depicted in Figure 3 for the optimal observer gain  $L = (0.1272, -0.0160)^T$ . For Case 2, let  $\mathcal{A} = 0.05$ ,  $p = 1 - \mathcal{A} + a_p$ ,  $a_p = \{0.01, 0.03\}$ , and  $L$  as in (27); then, the corresponding  $\bar{v}_p$  are  $\bar{v}_p = \{2.8970, 3.5208\}$  and the  $\underline{\epsilon}_p$ , computed through (23), are given by  $\underline{\epsilon}_p = \{21.16, 46.16\}$ . In Figure 4, we show the ellipsoidal bounds  $\tilde{\mathcal{E}}_\alpha^p$  on the reachable sets  $\tilde{\mathcal{R}}_\alpha^p$  obtained using Corollary 3.

**Remark 3** *Many numerical results considering hidden attacks with different distributions are presented in the accompanying paper [27] (Section 4). Also, extensive Monte-Carlo simulations showing the tightness of the bounds presented here are given in [27].*

## V. CONCLUSION

In this paper, for a class of discrete-time LTI systems subject to sensor/actuator noise, we have provided tools for *quantifying* and *minimizing* the negative impact of sensor attacks on the estimation error dynamics performance given how the opponent accesses the dynamics (i.e., through the controller by tampering with sensor measurements). We have proposed to use the *reachable set* as a measure of the impact of an attack given a chosen detection method. For given system dynamics and attack detection scheme, we have derived ellipsoidal bounds on these reachable sets using LMIs. Then, we have provided synthesis tools for minimizing these bounds (minimizing thus the reachable sets) by properly redesigning the detectors.

## REFERENCES

- [1] A. Cárdenas, S. Amin, Z. Lin, Y. Huang, C. Huang, and S. Sastry, "Attacks against process control systems: Risk assessment, detection, and response," in *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*, 2011, pp. 355–366.
- [2] F. Pasqualetti, F. Dorfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 58, pp. 2715–2729, 2013.
- [3] Y. Mo, E. Garone, A. Casavola, and B. Sinopoli, "False data injection attacks against state estimation in wireless sensor networks," in *Decision and Control (CDC), 2010 49th IEEE Conference on*, 2010, pp. 5967–5972.
- [4] C. Kwon, W. Liu, and I. Hwang, "Security analysis for cyber-physical systems against stealthy deception attacks," in *American Control Conference (ACC), 2013*, 2013, pp. 3344–3349.
- [5] F. Miao, Q. Zhu, M. Pajic, and G. J. Pappas, "Coding sensor outputs for injection attacks detection," in *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, 2014, pp. 5776–5781.
- [6] C. M. Ahmed, C. Murguía, and J. Ruths, "Model-based attack detection scheme for smart water distribution networks," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, ser. ASIA CCS '17, 2017, pp. 101–113.
- [7] E. Rothstein Morris, C. Murguía, and M. Ochoa, "Design-time quantification of integrity in cyber-physical-systems," in *Proceedings of the 2017 ACM SIGSAC Workshop on Programming Languages and Analysis for Security*, (accepted), 2017.
- [8] C. Z. Bai, F. Pasqualetti, and V. Gupta, "Security in stochastic control systems: Fundamental limitations and performance bounds," in *American Control Conference (ACC), 2015*, 2015, pp. 195–200.
- [9] C. Scherer and S. Weiland, *Linear Matrix Inequalities in Control*. The Netherlands: Springer-Verlag, 2000.
- [10] J. Chen and R. J. Patton, *Robust Model-based Fault Diagnosis for Dynamic Systems*. Norwell, MA, USA: Kluwer Academic Publishers, 1999.
- [11] E. Kyriakides and M. M. Polycarpou, Eds., *Intelligent Monitoring, Control, and Security of Critical Infrastructure Systems*, ser. Studies in Computational Intelligence. Springer, 2015, vol. 565.
- [12] C. Murguía and J. Ruths, "Characterization of a cusum model-based sensor attack detector," in *proceedings of the 55th IEEE Conference on Decision and Control (CDC)*, 2016.
- [13] —, "Cusum and chi-squared attack detection of compromised sensors," in *proceedings of the IEEE Multi-Conference on Systems and Control (MSC)*, 2016.
- [14] J. Giraldo, A. Cardenas, and N. Quijano, "Integrity attacks on real-time pricing in smart grids: Impact and countermeasures," *IEEE Transactions on Smart Grid*, vol. PP, 2016.
- [15] A. Cardenas, S. Amin, B. Sinopoli, A. Giani, A. Perrig, and S. Sastry, "Challenges for securing cyber physical systems," in *Workshop on Future Directions in Cyber-physical Systems Security*, 2009.
- [16] Z. Guo, D. Shi, K. H. Johansson, and L. Shi, "Optimal Linear Cyber-Attack on Remote State Estimation," *IEEE Transactions on Control of Network Systems*, vol. PP, no. 99, pp. 1–10, 2016.
- [17] Y. Mo and B. Sinopoli, "On the performance degradation of cyber-physical systems under stealthy integrity attacks," *IEEE Transactions on Automatic Control*, vol. 61, pp. 2618–2624, 2016.
- [18] C. Murguía, N. van de Wouw, and J. Ruths, "Reachable sets of hidden cps sensor attacks: Analysis and synthesis tools," in *proceedings of the IFAC World Congress*, 2016.
- [19] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, ser. Studies in Applied Mathematics. Philadelphia, PA: SIAM, 1994, vol. 15.
- [20] D. Luenberger, *Introduction to dynamic systems : theory, models, and applications*. New York: Wiley, 1979.
- [21] K. J. Aström and B. Wittenmark, *Computer-controlled Systems (3rd Ed.)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1997.
- [22] M. Ross, *Introduction to Probability Models, Ninth Edition*. Orlando, FL, USA: Academic Press, Inc., 2006.
- [23] C. van Dobben de Bruyn, *Cumulative sum tests : theory and practice*. London : Griffin, 1968.
- [24] B. Adams, W. Woodall, and C. Lowry, "The use (and misuse) of false alarm probabilities in control chart design," *Frontiers in Statistical Quality Control 4*, pp. 155–168, 1992.
- [25] N. D. That, P. T. Nam, and Q. P. Ha, "Reachable set bounding for linear discrete-time systems with delays and bounded disturbances," *Journal of Optimization Theory and Applications*, vol. 157, pp. 96–107, 2013.
- [26] J. Lofberg, "Yalmip : a toolbox for modeling and optimization in matlab," in *Computer Aided Control Systems Design, 2004 IEEE International Symposium on*, 2004, pp. 284–289.
- [27] N. Hashemil, C. Murguía, and J. Ruths, "A comparison of stealthy sensor attacks on control systems," in eprint arXiv (submitted to ACC2018), 2017.