Secure State Estimation Against Sensor Attacks in the Presence of Noise

Shaunak Mishra, Yasser Shoukry, Nikhil Karamchandani, Suhas N. Diggavi, *Fellow, IEEE*, and Paulo Tabuada

***Abstract*—We consider the problem of estimating the state of a noisy linear dynamical system when an unknown subset of sensors is arbitrarily corrupted by an adversary. We propose a secure state estimation algorithm, and derive (optimal) bounds on the achievable state estimation error given an upper bound on the number of attacked sensors. The proposed state estimator involves Kalman filters operating over subsets of sensors to search for a sensor subset which is reliable for state estimation. To further improve the subset search time, we propose Satisfiability Modulo Theory-based techniques to exploit the combinatorial nature of searching over sensor subsets. Finally, as a result of independent interest, we give a coding theoretic view of attack detection and state estimation against sensor attacks in a noiseless dynamical system.**

***Index Terms*—Secure cyber-physical systems, secure state reconstruction, sensor attacks.**

I. Introduction

SECURING cyberphysical systems (CPS) is a problem of growing importance as the vast majority of today's critical infrastructure is managed by such systems. In this context, it is crucial to understand the fundamental limits for state estimation, an integral aspect of CPS, in the presence of malicious attacks. With this motivation, we focus on securely estimating the state of a linear dynamical system from a set of noisy and maliciously corrupted sensor measurements. We restrict the sensor attacks to be sparse in nature, that is, an adversary can arbitrarily corrupt an unknown subset of sensors in the system but is restricted by an upper bound on the number of attacked sensors.

Several recent works have studied the problem of secure state estimation against sensor attacks in linear dynamical systems. For setups with no noise in sensor measurements, the results reported in [2]–[5] show that given a strong notion of observability, (sparse) sensor attacks can always be detected and isolated, and we can exactly estimate the state of the system. However, with noisy sensors, it is not trivial to distinguish between the noise and the attacks injected by an adversary. Prior work on state estimation with sensor attacks in the presence of noise can be broadly divided into two categories depending on the noise model: 1) bounded nonstochastic noise and 2) Gaussian noise. The results reported in [6]–[8] deal with bounded nonstochastic noise. Though they provide sufficient conditions for distinguishing the sparse attack vector from bounded noise, they do not guarantee the optimality of their estimation algorithm. The problem we focus on in this paper falls into the second category, that is, sensor attacks in the presence of Gaussian noise. Prior work in this category includes [9]–[12]. In [9], the focus is on detecting a class of sensor attacks called *replay* attacks where the attacker replaces legitimate sensor outputs with outputs from previous time instants. In [10], the performance degradation of a scalar Kalman filter (that is, scalar state and a single sensor) is studied when the (single) sensor is under attack. They do not study attack sparsity across multiple sensors and, in addition, they focus on an adversary whose objective is to degrade the estimation performance without being detected (leading to a restricted class of sensor attacks). In [11] and [12], robustification approaches for state estimation against sparse sensor attacks are studied. However, they lack optimality guarantees against arbitrary sensor attacks.

In this paper, we study a general linear dynamical system with process and sensor noises that have a Gaussian distribution, and give (optimal) guarantees on the achievable state estimation error against arbitrary sensor attacks. The following toy example is illustrative of the nature of the problem addressed in this paper and some of the ideas behind our solution.

Example 1: Consider a linear dynamical system with a scalar state $x(t)$ such that $x(t+1) = x(t) + w(t)$, and three sensors (indexed by $d \in \{1, 2, 3\}$) with outputs $y_d(t) = x(t) + v_d(t)$; where $w(t)$ and $v_d(t)$ are the process noise and sensor noise at sensor d, respectively. The process and sensor noises follow a zero mean Gaussian distribution with i.i.d. instantiations over time. The sensor noise is also independent across sensors. Now, consider an adversary which can attack any one of the sensors in the system and arbitrarily change its output. In the absence of sensor noise, it is trivial to detect such an attack since the two good sensors (not attacked by the adversary) will have the same output. Hence, a majority-based rule on the outputs leads to the exact state. However, in the presence of sensor noise, a difference in outputs across sensors can also be attributed to the noise and, thus, cannot be considered an attack indicator. As a

Manuscript received October 8, 2015; revised June 18, 2016; accepted August 22, 2016. Date of publication September 7, 2016; date of current version March 16, 2017. This work was supported by the National Science Foundation under Grant 1136174 and DARPA under agreement no. FA8750-12-2-0247. A preliminary version of this work appeared in the proceedings of ISIT 2015 [1]. Recommended by Associate Editor Bruno Sinopoli.

S. Mishra, S. N. Diggavi, and P. Tabuada are with the Electrical Engineering Department, University of California, Los Angeles, CA 90095-1594 USA (e-mail: shaunakmishra@ucla.edu; suhasdiggavi@ucla.edu; tabuada@ucla.edu).

Y. Shoukry is with both the Department of Electrical Engineering and Computer Science, University of California, Berkeley as well as the Electrical Engineering Department, University of California, Los Angeles (e-mail: yshoukry@eecs.berkeley.edu).

N. Karamchandani is with the Electrical Engineering Department, Indian Institute of Technology Bombay, Mumbai 400076 India (e-mail: nikhilk@ee.iitb.ac.in).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCNS.2016.2606880

consequence of results in this paper, in this example, we can identify a subset of two sensors which can be reliably used for state estimation despite an adversary who can attack any one of the three noisy sensors. In particular, our approach for this example would be to search for a subset of two sensors which satisfy the following check: over a large enough time window, the outputs from the two sensors are *consistent* with the Kalman state estimate based on outputs from the same subset of sensors. Furthermore, we can show that such an approach leads to the optimal state estimation error for the given adversarial setup.

In this paper, we generalize the Kalman filter-based approach in the aforementioned example to a general linear dynamical system with sensor and process noise. The Kalman estimate-based check mentioned in the aforementioned example forms the basis of a detector for an *effective* attack; a notion that we introduce in this paper. For state estimation, we search for a sensor subset which passes such an effective attack detector, and then use outputs from such a sensor subset for state estimation. We also derive impossibility results (lower bounds) on the state estimation error in our adversarial setup, and show that our proposed state estimation algorithm is optimal in the sense that it achieves these lower bounds. To further reduce the sensor subset search time for the state estimator, we propose Satisfiability Modulo Theory (SMT)-based techniques to harness the combinatorial nature of the search problem, and demonstrate the improvements in search time through numerical experiments.

As a result of independent interest, we give a coding-theoretic interpretation (alternate proof) for the necessary and sufficient conditions for secure state estimation in the absence of noise [3], [4], [7] (known as the sparse observability condition). In particular, we relate the sparse observability condition required for attack detection and secure state estimation in dynamical systems to the Hamming distance requirements for error detection and correction [13] in classical coding theory.

The remainder of this paper[1] is organized as follows. Section II deals with the setup and problem formulation. In Section III, we describe our effective attack detector followed by Section IV on our main results for effective attack detection and secure state estimation. Section V deals with SMT-based techniques and Section VI deals with the experimental results. Finally, Section VII describes the coding-theoretic view for attack detection and secure state estimation.

II. Setup

In this section, we discuss the adversarial setup along with assumptions on the underlying dynamical system, and provide a mathematical formulation of the state estimation problem considered in this paper.

A. Notation

Symbols $\mathbb{N}, \mathbb{R}$, and $\mathbb{B}$ denote the sets of natural, real, and Boolean numbers, respectively. The symbol $\wedge$ denotes the logical AND operator. The support of a vector $\mathbf{x} \in \mathbb{R}^n$, denoted by $\text{supp}(\mathbf{x})$, is the set of indices of the nonzero elements of $\mathbf{x}$. If $\mathbf{s}$ is a set, $|\mathbf{s}|$ is the cardinality of $\mathbf{s}$. For the matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$, unless stated otherwise, we denote by $\mathbf{M}_i \in \mathbb{R}^{1\times n}$ the ith row of the matrix. For the set $\mathbf{s} \subseteq \{1, \ldots, m\}$, we denote by $\mathbf{M}_{\mathbf{s}} \in \mathbb{R}^{|\mathbf{s}|\times n}$ the matrix obtained from $\mathbf{M}$ by removing all of the rows except those indexed by $\mathbf{s}$. We use $tr\left(\mathbf{M}\right)$ to denote the trace of the matrix $\mathbf{M}$. If the matrix $\mathbf{M}$ is symmetric, we use $\lambda_{\min}\left(\mathbf{M}\right)$ and $\lambda_{\max}\left(\mathbf{M}\right)$ to denote the minimum and maximum eigenvalue of $\mathbf{M}$, respectively. We denote by $\mathbb{S}^n_+$ the set of all $n\times n$ positive semidefinite matrices. For a random variable $\mathbf{x} \in \mathbb{R}^n$, we denote its mean by $\mathbb{E}\left(\mathbf{x}\right) \in \mathbb{R}$ and its covariance by $Var(\mathbf{x}) \in \mathbb{S}^n_+$. For a discrete time random process $\{\mathbf{x}(t)\}_{t\in\mathbb{N}}$, the sample average of $\mathbf{x}$ using N samples starting at time t_1 is defined as follows:

$$\mathbb{E}_{N,t_1}\left(\mathbf{x}\right) = \frac{1}{N}\sum_{t=t_1}^{t_1+N-1}\mathbf{x}(t). \tag{1}$$

We denote by $\mathbf{I}_m \in \mathbb{R}^{m\times m}$ and $\mathbf{1}_m \in \mathbb{R}^{m\times 1}$ the identity matrix of dimension m and the vector of all ones, respectively. The notation $\mathbf{x}(t) \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Omega}\right)$ is used to denote an i.i.d. Gaussian random process with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Omega}$. Finally, we use the symbol $\preccurlyeq$ for element-wise comparison between matrices. That is, for two matrices $\mathbf{A}$ and $\mathbf{B}$ of the same size, $\mathbf{A} \preccurlyeq \mathbf{B}$ is true if and only if each element $a_{i,j}$ is smaller than or equal to $b_{i,j}$.

B. System Model

We consider a linear dynamical system $\boldsymbol{\Sigma}_a$ with sensor attacks as shown

$$\boldsymbol{\Sigma}_a \begin{cases} \mathbf{x}\left(t+1\right) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) + \mathbf{w}(t), \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t) + \mathbf{v}(t) + \mathbf{a}(t), \end{cases} \tag{2}$$

where $\mathbf{x}(t) \in \mathbb{R}^n$ denotes the state of the plant at time $t \in \mathbb{N}$, $\mathbf{u}(t) \in \mathbb{R}^m$ denotes the input at time t, $\mathbf{w}(t) \sim \mathcal{N}\left(\mathbf{0}, \sigma_w^2\mathbf{I}_n\right)$ denotes the process noise at time t, $\mathbf{y}(t) \in \mathbb{R}^p$ denotes the output of the plant at time t, and $\mathbf{v}(t) \sim \mathcal{N}\left(\mathbf{0}, \sigma_v^2\mathbf{I}_p\right)$ denotes the sensor noise at time t. Both $\mathbf{v}(t)$ and $\mathbf{w}(t)$ have i.i.d. instantiations over time, and $\mathbf{v}(t)$ is independent of $\mathbf{w}(t)$. In addition, we denote the output and (sensor) noise at sensor $i \in \{1, 2, \ldots, p\}$ at time t as $y_i(t) \in \mathbb{R}$ and $v_i(t) \in \mathbb{R}$, respectively. We assume that the input $\mathbf{u}(t)$ is known at all times. Hence, its contribution to the output $\mathbf{y}(t)$ is also known and, therefore, $\mathbf{u}(t)$ can be ignored. That is, for the rest of this paper and without loss of generality, we consider the case of $\mathbf{u}(t) = 0$ for all time $t \in \mathbb{N}$.

The sensor attack vector $\mathbf{a}(t) \in \mathbb{R}^p$ in (2) is introduced by a k-adversary defined as follows.

Assumption 1: A k-adversary can corrupt any k out of the p sensors in the system.

Specifically, let $\boldsymbol{\kappa} \subseteq \{1, 2, \ldots, p\}$ denote the set of attacked sensors (with $|\boldsymbol{\kappa}| = k$). The k-adversary can observe the actual outputs in the k attacked sensors and change them arbitrarily. For an attack-free sensor $j \notin \boldsymbol{\kappa}$, $\mathbf{a}_j(t) = 0$ for all time $t \in \mathbb{N}$.

Assumption 2: The adversary's choice of $\boldsymbol{\kappa}$ is unknown but is assumed to be constant over time (static adversary).

Assumption 3: The adversary is assumed to have unbounded computational power and knows the system parameters (*e.g.*, $\mathbf{A}$ and $\mathbf{C}$) and noise statistics (*e.g.*, σ_w^2 and σ_v^2).

[1] Compared to the preliminary version [1], this paper differs in the presentation of results through effective attack detection. In addition, we reduce the complexity of the state estimation algorithm in [1] and describe SMT-based techniques for reducing the subset search time.

However, the adversary is limited to have only causal knowledge of the process and sensor noise as stated by the following two assumptions.

Assumption 4: The adversary's knowledge at time t is statistically independent of $\mathbf{w}(t')$ for $t' > t$, i.e., $\mathbf{a}(t)$ is statistically independent of $\{\mathbf{w}(t')\}_{t'>t}$.

Assumption 5: For an attack-free sensor $i \in \{1, 2, \ldots, p\} \setminus \boldsymbol{\kappa}$, the adversary's knowledge at time t (and, hence, $\mathbf{a}(t)$) is statistically independent of $\{v_i(t')\}_{t'>t}$.

Intuitively, Assumptions 4 and 5 limit the adversary to have only causal knowledge of the process noise and the sensor noise in *good* sensors (not attacked by the adversary). Note that, apart from Assumptions 4 and 5, we do not impose any restrictions on the statistical properties, boundedness, and the time evolution of the corruptions introduced by the k-adversary.

In the following subsections, we first introduce the (effective) attack detection problem, followed by the (optimal) secure state estimation problem. As we show later in this paper (in Section IV), our solution for the effective attack detection problem is used as a crucial component for solving the secure state estimation problem.

C. Effective Attack Detection Problem

In this section, we introduce our notion of effective (sensor) attacks and formulate the problem of detecting them. Recall that in the absence of sensor attacks, using a Kalman filter for estimating the state in (2) leads to the (optimal) minimum mean square error (MMSE) covariance asymptotically [14]. In this context, our notion of effective attacks is based on the following intuition: if we naively use a Kalman filter for state estimation in the presence of an adversary, an attack is effective when it causes a higher empirical error variance compared to the attack-free case. Before we formally state our definition of effective attacks, we first setup some notation for Kalman filters as described below.

We denote by $\hat{\mathbf{x}}_{\mathbf{s}}(t)$ the state estimate of a Kalman filter at time t using outputs till time $t-1$ from the sensor subset $\mathbf{s} \subseteq \{1, 2, \ldots, p\}$. Since we use outputs until time $t-1$, we essentially use the *prediction* version of a Kalman filter as opposed to *filtering* where outputs until time t are used to compute $\hat{\mathbf{x}}_{\mathbf{s}}(t)$. In this paper, we state our results using the prediction version of the Kalman filter; the extension for the filtering version is straightforward. (For details about the filtering version of our results, see the extended version [15].) In addition to $\hat{\mathbf{x}}_{\mathbf{s}}(t)$, we denote $\hat{\mathbf{x}}^{\star}_{\mathbf{s}}(t)$ as the Kalman filter state estimate at time t using sensor subset $\mathbf{s}$ when all of the sensors in $\mathbf{s}$ are attack-free. We eliminate the subscript $\mathbf{s}$ from the previous notation whenever the Kalman filter uses all sensor measurements, that is, when $\mathbf{s} = \{1, \ldots, p\}$. In this paper, for the sake of simplicity, we assume that all Kalman filters we consider (in our proposed algorithms and their analysis) are in steady state [14] when they use uncorrupted sensor outputs. Hence, in the absence of attacks, the error covariance matrix $\mathbf{P}^{\star}(t) \in \mathbb{S}_n^+$ defined as

$$\mathbf{P}^{\star}(t) = \mathbf{P}^{\star} = \mathbb{E}\Big(\left(\mathbf{x}(t) - \hat{\mathbf{x}}^{\star}(t)\right) \left(\mathbf{x}(t) - \hat{\mathbf{x}}^{\star}(t)\right)^T \Big),$$

does not depend on time. In a similar spirit, we define the error covariance matrix $\mathbf{P}^{\star}_{\mathbf{s}} \in \mathbb{S}_n^+$ corresponding to a sensor subset $\mathbf{s} \subseteq \{1, 2, \ldots, p\}$ as

$$\mathbf{P}^{\star}_{\mathbf{s}} = \mathbb{E}(\mathbf{x}(t) - \hat{\mathbf{x}}^{\star}_{\mathbf{s}}(t))(\mathbf{x}(t) - \hat{\mathbf{x}}^{\star}_{\mathbf{s}}(t))^T.$$

Note that the error covariance matrix depends on the set of sensors involved in estimating the state. Also, the steady-state error has zero mean, that is, $\mathbb{E}\left(\mathbf{x}(t) - \hat{\mathbf{x}}^{\star}_{\mathbf{s}}(t)\right) = 0$. Using the above notation, we define an $(\epsilon, \mathbf{s})$-effective attack as follows.

Definition 1 ($(\epsilon, \mathbf{s})$-Effective Attack): Consider the linear dynamical system under attack $\boldsymbol{\Sigma_a}$ as defined in (2), and a k-adversary satisfying Assumptions 1–5. For the set of sensors $\mathbf{s}$, an $\epsilon > 0$, and a large enough $N \in \mathbb{N}$, an attack signal is called $(\epsilon, \mathbf{s})$-effective at time t_1 if the following bound holds:

$$tr\left(\mathbb{E}_{N,t_1}\left(\mathbf{e}_{\mathbf{s}}\mathbf{e}_{\mathbf{s}}^T\right)\right) > tr(\mathbf{P}^{\star}_{\mathbf{s}}) + \epsilon,$$

where $\mathbf{e}_{\mathbf{s}}(t) = \mathbf{x}(t) - \hat{\mathbf{x}}_{\mathbf{s}}(t)$, and $\mathbb{E}_{N,t_1}(\cdot)$ denotes the sample average as defined (1).

In other words, an attack is called $(\epsilon, \mathbf{s})$-effective if it can lead to a higher estimation error compared to the optimal estimation error in the absence of sensor attacks, using the same set of sensors $\mathbf{s}$. An attack is called $(\epsilon, \mathbf{s})$-ineffective if it is not $(\epsilon, \mathbf{s})$-effective. Essentially, we use $\mathbb{E}_{N,t_1}\left(\mathbf{e}_{\mathbf{s}}\mathbf{e}_{\mathbf{s}}^T\right)$ as a *proxy* for the state estimation error covariance matrix in the presence of attacks; a sample average is used instead of an expectation because the resulting error in the presence of attacks may not be ergodic. Also, since $\hat{\mathbf{x}}_{\mathbf{s}}(t)$ is computed using all measurements from time 0 until time $t-1$, Definition 1 implicitly takes into consideration the effect of attack signal $\mathbf{a}(t)$ for the time window starting from 0 until time $t_1 + N - 1$.

Using the above notion of an $(\epsilon, \mathbf{s})$-effective attack, we define the ϵ-effective attack detection problem as follows.

Problem 1. [ϵ-Effective Attack Detection Problem]: Consider the linear dynamical system under attack $\boldsymbol{\Sigma_a}$ as defined in (2), and a k-adversary satisfying Assumptions 1–5. Let $\mathbf{s}_{\text{all}}$ be the set of all sensors, that is, $\mathbf{s}_{\text{all}} = \{1, \ldots, p\}$. Given an $\epsilon > 0$, construct an attack indicator $\hat{d}_{\text{attack}} \in \{0, 1\}$ such that:

$$\hat{d}_{\text{attack}}(t_1) = \begin{cases} 1 & \textbf{if } \text{the attack is } (\epsilon, \mathbf{s}_{\text{all}})\text{-effective at time } t_1 \\ 0 & \textbf{otherwise}. \end{cases}$$

D. Optimal Secure State Estimation Problem

We now focus on the problem of estimating the state from the adversarially corrupted sensors. We start by showing a negative result stating that a certain estimation error bound may be impossible to achieve in the presence of a k-adversary. To do so, we define the sensor set that contains $p-k$ sensors and corresponds to the worst case Kalman estimate as

$$\mathbf{s}_{\text{worst},p-k} = \arg \max_{\substack{\mathbf{s} \subseteq \{1,2,\ldots,p\}, \\ |\mathbf{s}| = p-k}} tr(\mathbf{P}^{\star}_{\mathbf{s}}). \qquad (3)$$

The impossibility result can now be stated as follows.

Theorem 1 (Impossibility): Consider the linear dynamical system under attack $\boldsymbol{\Sigma_a}$ as defined in (2), and an oracle MMSE estimator that has knowledge of $\boldsymbol{\kappa}$, that is, the set of sensors attacked by a k-adversary. Then, there exists a choice of sensors

$\boldsymbol{\kappa}$ and an attack sequence $\mathbf{a}(t)$ such that the trace of the error covariance of the oracle estimator is bounded from below as follows:

$$tr\Big(\mathbb{E}\left(\mathbf{e}(t)\mathbf{e}^T(t)\right)\Big) \geq tr\Big(\mathbf{P}^{\star}_{\mathbf{s}_{\text{worst},\,p-k}}\Big), \tag{4}$$

where $\mathbf{e}(t)$ above is the oracle estimator's error.

Proof: Consider the attack scenario where the outputs from all attacked sensors are equal to zero, *that is*, the corruption $\mathbf{a}_j(t) = -\mathbf{C}_j\mathbf{x}(t) - v_j(t),\ \forall j \in \boldsymbol{\kappa}$. In such a scenario, the information collected from the attacked sensors cannot enhance the estimation performance, and the oracle estimator can simply use the remaining (attack free) sensors to achieve the best possible error performance. Hence, the result follows by picking $\boldsymbol{\kappa}$ such that $\boldsymbol{\kappa} = \{1, \ldots, p\} \setminus \mathbf{s}_{\text{worst},p-k}$. ■

In the context of Theorem 1, we define a state estimate to be optimal if it is guaranteed to achieve the lower bound shown in (4). This can be formalized as follows.

Problem 2. [Optimal Secure State Estimation Problem]: Consider the linear dynamical system under attack $\boldsymbol{\Sigma_a}$ as defined in (2), and a k-adversary satisfying Assumptions 1–5. For a time window $G = \{t_1, t_1+1, \ldots, t_1+N-1\}$, construct the state estimates $\{\hat{\mathbf{x}}(t)\}_{t\in G}$ such that

$$tr\Big(\mathbb{E}_{N,t_1}\left(\mathbf{e}\mathbf{e}^T\right)\Big) \leq tr\Big(\mathbf{P}^{\star}_{\mathbf{s}_{\text{worst},\,p-k}}\Big),$$

where $\mathbf{e}(t) = \mathbf{x}(t) - \hat{\mathbf{x}}(t)$ is the state estimation error.

Similar to Definition 1, we use the sample average $\mathbb{E}_{N,t_1}\left(\mathbf{e}\mathbf{e}^T\right)$ in Problem 2 (and not expectation) since the resulting error in the presence of attacks may not be ergodic.

III. Sparse Observability and $(\epsilon, \mathbf{s})$-Effective Attack Detection

In this section, we first describe the notion of k-sparse observability [4]. This notion plays an important role in determining when Problems 1 and 2 are solvable. After describing sparse observability, we describe an algorithm for $(\epsilon, \mathbf{s})$-effective attack detection which leverages sparse observability for its performance guarantees.

A. *k-Sparse Observability*

Definition 2. (k-Sparse Observable System): The linear dynamical system under attack $\boldsymbol{\Sigma_a}$ as defined in (2), is said to be k-sparse observable if for every set $\mathbf{s} \subseteq \{1, \ldots, p\}$ with $|\mathbf{s}| = p-k$, the pair $(A, C_{\mathbf{s}})$ is observable.

In other words, a system is k-sparse observable if it remains observable after eliminating any choice of k sensors. In the absence of sensor and process noise, the conditions under which exact (that is, zero error) state estimation can be done despite sensor attacks have been studied in [3], [4], and [7] where it is shown that $2k$-sparse observability is necessary and sufficient for exact state estimation against a k-adversary. In Section VII, we provide a coding-theoretic interpretation for this condition in the context of attack detection and secure state estimation in any noiseless dynamical system.

B. *$(\epsilon, \mathbf{s})$-Effective Attack Detector*

In this section, we describe an algorithm based on the sparse observability condition for detecting an $(\epsilon, \mathbf{s})$-effective attack. We first introduce some additional notations, followed by the description of the algorithm and its performance guarantee.

1) Additional Notation: Let the sensors be indexed by $i \in \{1, 2, \ldots, p\}$. We define the following observability matrices:

$$\mathcal{O}_i = \begin{bmatrix} \mathbf{C}_i \\ \mathbf{C}_i\mathbf{A} \\ \vdots \\ \mathbf{C}_i\mathbf{A}^{\mu_i-1} \end{bmatrix}, \quad \mathcal{O} = \begin{bmatrix} \mathcal{O}_1 \\ \mathcal{O}_2 \\ \vdots \\ \mathcal{O}_p \end{bmatrix}, \tag{5}$$

where $\mathcal{O}_i$ is the observability matrix for sensor i (with observability index μ_i as shown in (5)) and $\mathcal{O}$ is the observability matrix for the entire system (that is, p sensors) formed by stacking the observability matrices for the sensors. Similarly, for any sensor subset $\mathbf{s} \subseteq \{1, 2, \ldots, p\}$, we denote the observability matrix for $\mathbf{s}$ by $\mathcal{O}_{\mathbf{s}}$ (formed by stacking the observability matrices of sensors in $\mathbf{s}$). Without loss of generality, we will consider the observability index $\mu_i = n$ for each sensor. For any sensor subset $\mathbf{s}$ with $|\mathbf{s}| > k$, we define $\lambda_{\min,\mathbf{s}\setminus k}$ as follows:

$$\lambda_{\min,\mathbf{s}\setminus k} = \min_{\mathbf{s}_1 \subset \mathbf{s},\ |\mathbf{s}_1| = |\mathbf{s}|-k} \lambda_{\min}\left(\mathcal{O}_{\mathbf{s}_1}^T \mathcal{O}_{\mathbf{s}_1}\right), \tag{6}$$

where $\lambda_{\min}\left(\mathcal{O}_{\mathbf{s}_1}^T \mathcal{O}_{\mathbf{s}_1}\right)$ denotes the minimum eigenvalue of $\mathcal{O}_{\mathbf{s}_1}^T \mathcal{O}_{\mathbf{s}_1}$. We define matrices $\mathbf{J}_i$, J, and $\mathbf{M}$ as shown below

$$\mathbf{J}_i = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{C}_i & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{C}_i\mathbf{A} & \mathbf{C}_i & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_i\mathbf{A}^{\mu_i-2} & \mathbf{C}_i\mathbf{A}^{\mu_i-3} & \ldots & \mathbf{C}_i \end{bmatrix}, \quad J = \begin{bmatrix} \mathbf{J}_1 \\ \mathbf{J}_2 \\ \vdots \\ \mathbf{J}_p \end{bmatrix},$$

$$\mathbf{M} = \sigma_w^2 JJ^T + \sigma_v^2 \mathbf{I}_{np}. \tag{7}$$

In a similar spirit, $J_{\mathbf{s}}$ is defined for a sensor subset $\mathbf{s}$ by stacking $\mathbf{J}_i$ for $i \in \mathbf{s}$, and $\mathbf{M}_{\mathbf{s}} = \sigma_w^2 J_{\mathbf{s}} J_{\mathbf{s}}^T + \sigma_v^2 \mathbf{I}_{n|\mathbf{s}|}$. We use the following notation for sensor outputs and noises corresponding to a time window of size $\mu_i = n$ (observability index):

$$\mathbf{y}_i(t) = \begin{bmatrix} y_i(t) \\ y_i(t+1) \\ \vdots \\ y_i(t+\mu_i-1) \end{bmatrix}, \quad \mathbf{v}_i(t) = \begin{bmatrix} v_i(t) \\ v_i(t+1) \\ \vdots \\ v_i(t+\mu_i-1) \end{bmatrix},$$

$$\bar{\mathbf{y}}(t) = \begin{bmatrix} \mathbf{y}_1(t) \\ \mathbf{y}_2(t) \\ \vdots \\ \mathbf{y}_p(t) \end{bmatrix}, \quad \bar{\mathbf{v}}(t) = \begin{bmatrix} \mathbf{v}_1(t) \\ \mathbf{v}_2(t) \\ \vdots \\ \mathbf{v}_p(t) \end{bmatrix}, \quad \bar{\mathbf{w}}(t) = \begin{bmatrix} \mathbf{w}(t) \\ \mathbf{w}(t+1) \\ \vdots \\ \mathbf{w}(t+n-2) \end{bmatrix} \tag{8}$$

where $y_i(t)$ and $v_i(t)$ denote the output and sensor noise at sensor i at time t, respectively.

2) Attack Detection Algorithm: We consider the attack detection problem for a time window $G = \{t_1, t_1+1, \ldots, t_1+N-1\}$, and assume without loss of generality that the window

Algorithm 1: ATTACK-DETECT $(\mathbf{s}, t_1)$.

```
Run a Kalman filter that uses all measurements from sensors indexed by s until time $t_1 - 1$ and compute the estimate $\hat{\mathbf{x}}_\mathbf{s}(t_1) \in \mathbb{R}^n$.
Recursively repeat the previous step $N-1$ times to calculate all estimates $\hat{\mathbf{x}}_\mathbf{s}(t) \in \mathbb{R}^n, \forall t \in G = \{t_1, t_1+1, \ldots, t_1+N-1\}$.
For time $t \in G$, calculate the *block* residue
    $\mathbf{r}_\mathbf{s}(t) = \bar{\mathbf{y}}_\mathbf{s}(t) - \mathcal{O}_\mathbf{s}\hat{\mathbf{x}}_\mathbf{s}(t) \quad \forall t \in G.$
**if** block residue test defined below holds,
    $\mathbb{E}_{N,t_1}\left(\mathbf{r}_\mathbf{s}\mathbf{r}_\mathbf{s}^T\right) - \left(\mathcal{O}_\mathbf{s}\mathbf{P}_\mathbf{s}^\star\mathcal{O}_\mathbf{s}^T + \mathbf{M}_\mathbf{s}\right) \preccurlyeq \eta\, \mathbf{1}_{n|\mathbf{s}|}\mathbf{1}_{n|\mathbf{s}|}^T,$
    where $0 < \eta \leq \left(\frac{\lambda_{\min,\mathbf{s}\setminus k}}{3n(|\mathbf{s}|-k)}\right)\epsilon$, **then**
    assert $\hat{d}_{\text{attack},\mathbf{s}}(t_1) := 0$
**else**
    assert $\hat{d}_{\text{attack},\mathbf{s}}(t_1) := 1$
**end if**
**return** $(\hat{d}_{\text{attack},\mathbf{s}}(t_1), \{\hat{\mathbf{x}}_\mathbf{s}(t)\}_{t\in G})$.
```

size N is divisible by n. For a sensor subset $\mathbf{s}$ with $|\mathbf{s}| > k$, we start by computing the state estimate $\hat{\mathbf{x}}_\mathbf{s}(t_1)$ obtained through a Kalman filter that uses measurements collected from time 0 up to time $t_1 - 1$ from all sensors indexed by the subset $\mathbf{s}$. Using this estimate, we can calculate the *block* residue $\mathbf{r}_\mathbf{s}(t_1)$ which is the discrepancy between the estimated output $\hat{\bar{\mathbf{y}}}_\mathbf{s}(t_1) = \mathcal{O}_\mathbf{s}\hat{\mathbf{x}}_\mathbf{s}(t_1)$ and the actual output $\bar{\mathbf{y}}_\mathbf{s}(t_1)$, *i.e.*,

$$\mathbf{r}_\mathbf{s}(t_1) = \bar{\mathbf{y}}_\mathbf{s}(t_1) - \hat{\bar{\mathbf{y}}}_\mathbf{s}(t_1) = \bar{\mathbf{y}}_\mathbf{s}(t_1) - \mathcal{O}_\mathbf{s}\hat{\mathbf{x}}_\mathbf{s}(t_1). \tag{9}$$

By repeating the previous procedure $N-1$ times, we can obtain the sequence of residues $\{\mathbf{r}_\mathbf{s}(t)\}_{t\in G}$. The next step is to calculate the sample average of $\mathbf{r}_\mathbf{s}(t)\mathbf{r}_\mathbf{s}^T(t)$, and compare the sample average with the expected value of $\mathbf{r}_\mathbf{s}(t)\mathbf{r}_\mathbf{s}^T(t)$ in the case when sensor subset $\mathbf{s}$ is attack free. This can be done using the following (block) residue test:

$$\mathbb{E}_{N,t_1}\left(\mathbf{r}_\mathbf{s}\mathbf{r}_\mathbf{s}^T\right) - \left(\mathcal{O}_\mathbf{s}\mathbf{P}_\mathbf{s}^\star\mathcal{O}_\mathbf{s}^T + \mathbf{M}_\mathbf{s}\right) \preccurlyeq \eta\, \mathbf{1}_{n|\mathbf{s}|}\mathbf{1}_{n|\mathbf{s}|}^T, \tag{10}$$

for some $\eta > 0$. Simply put, the residue test checks whether the sample average of $\mathbf{r}_\mathbf{s}(t)\mathbf{r}_\mathbf{s}^T(t)$ over time window G is *close* to its attack-free expected value $\mathcal{O}_\mathbf{s}\mathbf{P}_\mathbf{s}^\star\mathcal{O}_\mathbf{s}^T + \mathbf{M}_\mathbf{s}$. This is similar in spirit to a Chi-squared test [16] (with stronger guarantees as shown in Section III-C), and the time window essentially *averages out* the effect of noise. Note the attack-free estimation error covariance matrix $\mathbf{P}_\mathbf{s}^\star$ used in (10) can be computed offline [14] without the need for any data collected from attack-free sensors. If the element-wise comparison in the residue test (10) is valid, we set the attack detection flag $\hat{d}_{\text{attack},\mathbf{s}}(t_1)$ to zero indicating that no attack was detected in sensor subset $\mathbf{s}$. This procedure is summarized in Algorithm 1.

C. Performance Guarantees

In this subsection, we describe our first main result which is concerned with the correctness of Algorithm 1.

Lemma 1: Let the linear dynamical system as defined in (2) be $2k$-sparse observable. Consider a k-adversary satisfying Assumptions 1–5 and a sensor subset $\mathbf{s} \subseteq \{1, 2, \ldots, p\}$ with $|\mathbf{s}| \geq p - k$. For any $\epsilon > 0$ and $\delta > 0$, there exists a large enough time window length N such that when Algorithm 1 terminates with $\hat{d}_{\text{attack},\mathbf{s}}(t_1) = 0$, the following probability bound holds:

$$\mathbb{P}\Big(tr\left(\mathbb{E}_{N,t_1}\left(\mathbf{e}_\mathbf{s}\mathbf{e}_\mathbf{s}^T\right) - \mathbf{P}_\mathbf{s}^\star\right) \leq \epsilon\Big) \geq 1 - \delta, \tag{11}$$

where $\mathbf{e}_\mathbf{s}(t) = \mathbf{x}(t) - \hat{\mathbf{x}}_\mathbf{s}(t)$. In other words, for large enough N, the bound $tr\left(\mathbb{E}_{N,t_1}\left(\mathbf{e}_\mathbf{s}\mathbf{e}_\mathbf{s}^T\right) - \mathbf{P}_\mathbf{s}^\star\right) \leq \epsilon$ holds with high probability[2] (w.h.p.). Moreover, in the context of $(\epsilon, \mathbf{s})$-effective attacks, the following also holds:

$$\mathbb{P}\left(\hat{d}_{\text{attack},\mathbf{s}}(t_1) = d_{\text{attack},\mathbf{s}}(t_1)\right) \geq 1 - \delta, \tag{12}$$

where $\hat{d}_{\text{attack},\mathbf{s}}(t_1)$ is the output of Algorithm 1 while $d_{\text{attack},\mathbf{s}}(t_1)$ is the output of an oracle detector that knows the exact set of attacked sensors. Hence, Algorithm 1 can detect any $(\epsilon, \mathbf{s})$-effective attack w.h.p. for large enough N.

Proof of Lemma 1: We focus only on showing that (11) holds whenever Algorithm 1 terminates with $\hat{d}_{\text{attack},\mathbf{s}}(t_1) = 0$; the rest of the lemma easily follows from the proof of (11) and Definition 1. Since we assume that the set $\mathbf{s}$ has cardinality $|\mathbf{s}| \geq p - k$, we can conclude that there exists a subset $\mathbf{s}_g \subset \mathbf{s}$ with cardinality $|\mathbf{s}_g| \geq p - 2k$ sensors such that all of its sensors are attack free (subscript g in $\mathbf{s}_g$ stands for *good* sensors in $\mathbf{s}$). Hence, by decomposing the set $\mathbf{s}$ into an attack-free set $\mathbf{s}_g$ and a potentially attacked set $\mathbf{s} \setminus \mathbf{s}_g$, we can conclude that after a permutation similarity transformation for (10), the following holds for the attack-free subset: $\mathbf{s}_g$:

$$\mathbb{E}_{N,t_1}\left(\mathbf{r}_{\mathbf{s}_g}\mathbf{r}_{\mathbf{s}_g}^T\right) - \mathcal{O}_{\mathbf{s}_g}\mathbf{P}_\mathbf{s}^\star\mathcal{O}_{\mathbf{s}_g}^T - \mathbf{M}_{\mathbf{s}_g} \preccurlyeq \eta\, \mathbf{1}_{n(|\mathbf{s}|-k)}\mathbf{1}_{n(|\mathbf{s}|-k)}^T.$$

Therefore

$$\begin{aligned} tr\left(\mathbb{E}_{N,t_1}\left(\mathbf{r}_{\mathbf{s}_g}\mathbf{r}_{\mathbf{s}_g}^T\right) - \mathcal{O}_{\mathbf{s}_g}\mathbf{P}_\mathbf{s}^\star\mathcal{O}_{\mathbf{s}_g}^T - \mathbf{M}_{\mathbf{s}_g}\right)& \\ \leq n(|\mathbf{s}|-k)\eta = \epsilon_1.& \end{aligned} \tag{13}$$

Similarly, after a suitable permutation Π, we can decompose the block residue $\mathbf{r}_\mathbf{s}(t)$ defined in (9) as follows:

$$\begin{aligned} \Pi\left(\mathbf{r}_\mathbf{s}(t)\right) &= \begin{bmatrix} \mathbf{r}_{\mathbf{s}_g}(t) \\ \mathbf{r}_{\mathbf{s}\setminus\mathbf{s}_g}(t) \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{y}}_{\mathbf{s}_g}(t) - \mathcal{O}_{\mathbf{s}_g}\hat{\mathbf{x}}_\mathbf{s}(t) \\ \bar{\mathbf{y}}_{\mathbf{s}\setminus\mathbf{s}_g}(t) - \mathcal{O}_{\mathbf{s}\setminus\mathbf{s}_g}\hat{\mathbf{x}}_\mathbf{s}(t) \end{bmatrix} \\ &= \begin{bmatrix} \mathcal{O}_{\mathbf{s}_g}\mathbf{x}(t) + J_{\mathbf{s}_g}\bar{\mathbf{w}}(t) + \bar{\mathbf{v}}_{\mathbf{s}_g}(t) - \mathcal{O}_{\mathbf{s}_g}\hat{\mathbf{x}}_\mathbf{s}(t) \\ \bar{\mathbf{y}}_{\mathbf{s}\setminus\mathbf{s}_g}(t) - \mathcal{O}_{\mathbf{s}\setminus\mathbf{s}_g}\hat{\mathbf{x}}_\mathbf{s}(t) \end{bmatrix} \\ &= \begin{bmatrix} \mathcal{O}_{\mathbf{s}_g}\mathbf{e}_\mathbf{s}(t) + \mathbf{z}_{\mathbf{s}_g}(t) \\ \bar{\mathbf{y}}_{\mathbf{s}\setminus\mathbf{s}_g}(t) - \mathcal{O}_{\mathbf{s}\setminus\mathbf{s}_g}\hat{\mathbf{x}}_\mathbf{s}(t) \end{bmatrix}, \end{aligned} \tag{14}$$

[2] By stating that the bound holds with high probability for large enough N, we mean that for any $\delta > 0$ and $\epsilon > 0$, $\exists N_{\delta,\epsilon} \in \mathbb{N}$ such that for $N > N_{\delta,\epsilon}$, $\mathbb{P}\Big(tr\left(\mathbb{E}_{N,t_1}\left(\mathbf{e}_\mathbf{s}\mathbf{e}_\mathbf{s}^T\right) - \mathbf{P}_\mathbf{s}^\star\right) \leq \epsilon\Big) \geq 1 - \delta$.

where $\mathbf{z}_{\mathbf{s}_g}(t) = J_{\mathbf{s}_g}\bar{\mathbf{w}}(t) + \bar{\mathbf{v}}_{\mathbf{s}_g}(t)$. Using (14), we can rewrite $tr\left(\mathbb{E}_{N,t_1}\left(\mathbf{r}_{\mathbf{s}_g}\mathbf{r}_{\mathbf{s}_g}^T\right)\right)$ as

$$\begin{aligned} tr\left(\mathbb{E}_{N,t_1}\left(\mathbf{r}_{\mathbf{s}_g}\mathbf{r}_{\mathbf{s}_g}^T\right)\right) &= tr\left(\mathcal{O}_{\mathbf{s}_g}\mathbb{E}_{N,t_1}\left(\mathbf{e}_{\mathbf{s}}\mathbf{e}_{\mathbf{s}}^T\right)\mathcal{O}_{\mathbf{s}_g}^T\right) \\ &+ tr\left(\mathbb{E}_{N,t_1}\left(\mathbf{z}_{\mathbf{s}_g}\mathbf{z}_{\mathbf{s}_g}^T\right)\right) + 2\mathbb{E}_{N,t_1}\left(\mathbf{e}_{\mathbf{s}}^T\mathcal{O}_{\mathbf{s}_g}^T\mathbf{z}_{\mathbf{s}_g}\right). \quad (15) \end{aligned}$$

By combining (13) and (15)

$$\begin{aligned} tr&\left(\mathcal{O}_{\mathbf{s}_g}\mathbb{E}_{N,t_1}\left(\mathbf{e}_{\mathbf{s}}\mathbf{e}_{\mathbf{s}}^T\right)\mathcal{O}_{\mathbf{s}_g}^T - \mathcal{O}_{\mathbf{s}_g}\mathbf{P}_{\mathbf{s}}^{\star}\mathcal{O}_{\mathbf{s}_g}^T\right) \\ &\leq tr\left(\mathbf{M}_{\mathbf{s}_g}\right) - tr\left(\mathbb{E}_{N,t_1}\left(\mathbf{z}_{\mathbf{s}_g}\mathbf{z}_{\mathbf{s}_g}^T\right)\right) + \epsilon_1 \\ &\quad - 2\mathbb{E}_{N,t_1}\left(\mathbf{e}_{\mathbf{s}}^T\mathcal{O}_{\mathbf{s}_g}^T\mathbf{z}_{\mathbf{s}_g}\right) \\ &\overset{(a)}{\leq} 2\epsilon_1 - 2\mathbb{E}_{N,t_1}\left(\mathbf{e}_{\mathbf{s}}^T\mathcal{O}_{\mathbf{s}_g}^T\mathbf{z}_{\mathbf{s}_g}\right) \quad (16) \\ &\overset{(b)}{\leq} 3\epsilon_1, \quad (17) \end{aligned}$$

where (a) follows w.h.p. due to the law of large numbers (LLN) for large enough N (details in Appendix A1), and (b) follows w.h.p. by showing that the cross term $2\mathbb{E}_{N,t_1}(\mathbf{e}^T\mathcal{O}_{\mathbf{s}_g}^T\mathbf{z}_{\mathbf{s}_g})$ has zero mean and vanishingly small variance for large enough N. The cross term analysis is described in detail in Appendix A2. Now recall that for any two matrices $\mathbf{A}$ and $\mathbf{B}$ of appropriate dimensions, $tr\,(\mathbf{AB}) = tr\,(\mathbf{BA})$. Using this fact along with (17), the following holds:

$$tr\left(\mathbb{E}_{N,t_1}\left(\mathbf{e}_{\mathbf{s}}\mathbf{e}_{\mathbf{s}}^T - \mathbf{P}_{\mathbf{s}}^{\star}\right)\mathcal{O}_{\mathbf{s}_g}^T\mathcal{O}_{\mathbf{s}_g}\right) \leq 3\epsilon_1, \quad (18)$$

and, hence, we obtain the following bound which completes the proof:

$$tr\left(\mathbb{E}_{N,t_1}\left(\mathbf{e}_{\mathbf{s}}\mathbf{e}_{\mathbf{s}}^T\right) - \mathbf{P}_{\mathbf{s}}^{\star}\right) \overset{(c)}{\leq} \frac{3\epsilon_1}{\lambda_{\min}\left(\mathcal{O}_{\mathbf{s}_g}^T\mathcal{O}_{\mathbf{s}_g}\right)} \overset{(d)}{\leq} \frac{3\epsilon_1}{\lambda_{\min,\mathbf{s}\backslash k}} \leq \epsilon \quad (19)$$

where (c) follows from Lemma 3 in Appendix B and (d) follows from the definition of $\lambda_{\min,\mathbf{s}\backslash k}$. Note that it follows from: $|\mathbf{s}_g| \geq p - 2k$ and $2k$-sparse observability, that $\lambda_{\min}(\mathcal{O}_{\mathbf{s}_g}^T\mathcal{O}_{\mathbf{s}_g})$ and $\lambda_{\min,\mathbf{s}\backslash k}$ are bounded away from zero. This completes the proof of (11). Based on the proof of (11) and Definition 1, it is now straightforward to show (12). Intuitively, the probability of mismatch between $\hat{d}_{\text{attack},\mathbf{s}}(t_1)$ and $d_{\text{attack},\mathbf{s}}(t_1)$ in (12) stems from the chances of a false detection; this occurs when the noise realizations deviate from LLN and the detector's threshold check fails despite the absence of an adversary. As seen in the proof of (11), w.h.p. the noise realizations obey LLN, and, hence, w.h.p. $\hat{d}_{\text{attack},\mathbf{s}}(t_1)$ and $d_{\text{attack},\mathbf{s}}(t_1)$ are equal. ■

IV. Effective Attack Detection and Secure State Estimation

Based on the performance guarantees for the Attack-Detect algorithm described in Section III, in this section, we describe our main results for Problems 1 and 2.

Algorithm 2: Exhaustive Search.

```
1: Enumerate all sets $\mathbf{s} \in \mathbf{S}$ such that
       $\mathbf{S} = \{\mathbf{s} | \mathbf{s} \subset \{1, 2, \ldots, p\},\ |\mathbf{s}| = p - k\}$.
2: Exhaustively search for $\mathbf{s}^* \in \mathbf{S}$ for which
   $d_{\text{attack},\mathbf{s}^*}(t_1) = 0$ and use $\hat{\mathbf{x}}_{\mathbf{s}^*}(t)$ for $t \in G$ as the state
   estimate.
```

A. Attack Detection

We start by showing a solution to Problem 1 (ϵ-effective attack detection), which follows directly from Lemma 1.

Theorem 2: Let the linear dynamical system defined in (2) be a k-sparse observable system. Consider a k-adversary satisfying Assumptions 1–5, and the detector $\hat{d}_{\text{attack}}(t_1) =$ Attack-Detect$(\mathbf{s}_{\text{all}}, t_1)$ where the set $\mathbf{s}_{\text{all}} = \{1, \ldots, p\}$. Then, for a large enough time window length N, w.h.p. $\hat{d}_{\text{attack}}(t_1)$ is equal to the attack indicator which solves Problem 1.

Proof: The proof is similar to the proof of Lemma 1. In the proof of Lemma 1, we basically required the set of good sensors $\mathbf{s}_g$ to form an observable system. Similarly, while checking for effective attacks on a sensor set of size p, we require the set of good sensors (of size $\geq p - k$) to form an observable system in order to repeat the steps in the proof for Lemma 1; this requirement is guaranteed by the k-sparse observability condition. On a related note, in Section VII, we give a coding-theoretic interpretation for the k-sparse observability requirement for attack detection. ■

B. Secure State Estimation

Algorithm 2 describes our proposed solution for Problem 2 (secure state estimation). As described in Algorithm 2, we exhaustively enumerate $\binom{p}{p-k}$ sensor subsets of size $p - k$, and then apply Attack-Detect on each sensor subset until we find one subset $\mathbf{s}^*$ for which Attack-Detect returns $\hat{d}_{\text{attack},\mathbf{s}^*}(t_1) = 0$ indicating that the subset is (ϵ-effective) attack free. The following theorem states the performance guarantees associated with Algorithm 2.

Theorem 3: Let the linear dynamical system defined in (2) be a $2k$-sparse observable system. Consider a k-adversary satisfying Assumptions 1–5. Consider the state estimate $\hat{\mathbf{x}}_{\mathbf{s}^*}(t)$ computed by Algorithm 2. Then, for any $\epsilon > 0$ and $\delta > 0$, there exists a large enough N such that

$$\mathbb{P}\left(tr\left(\mathbb{E}_{N,t_1}\left(\mathbf{e}_{\mathbf{s}^*}\mathbf{e}_{\mathbf{s}^*}^T\right)\right) \leq tr\left(\mathbf{P}_{\mathbf{s}_{\text{worst},p-k}}^{\star}\right) + \epsilon\right) \geq 1 - \delta \quad (20)$$

where $\mathbf{e}_{\mathbf{s}^*}(t) = \mathbf{x}(t) - \hat{\mathbf{x}}_{\mathbf{s}^*}(t)$ is the estimation error using $\hat{\mathbf{x}}_{\mathbf{s}^*}(t)$ as the state estimate. In other words, w.h.p. Algorithm 2 achieves the bound $\limsup_{N\to\infty} \frac{1}{N}\sum_{t\in G}\mathbf{e}_{\mathbf{s}^*}^T(t)\mathbf{e}_{\mathbf{s}^*}(t) \leq tr(\mathbf{P}_{\mathbf{s}_{\text{worst},p-k}}^{\star})$.

Proof: The result follows from Lemma 1 which ensures that in the absence of the $(\epsilon, \mathbf{s})$-effective attack property, the calculated state estimate still guarantees the bound (11). This, in turn, implies that in the worst case $\limsup_{N\to\infty} \frac{1}{N}\sum_{t\in G}\mathbf{e}_{\mathbf{s}^*}^T(t)\mathbf{e}_{\mathbf{s}^*}(t) = tr(\mathbf{P}_{\mathbf{s}_{\text{worst},p-k}}^{\star})$ is achievable.

However, since the k-adversary may not always attack the worst case set of sensors $\mathbf{s}_{\text{worst},p-k}$, we can replace the equality sign above with an inequality, leading to $\limsup_{N\to\infty} \frac{1}{N}\sum_{t\in G} \mathbf{e}^T_{\mathbf{s}^*}(t)\mathbf{e}_{\mathbf{s}^*}(t) \leq tr(\mathbf{P}^\star_{\mathbf{s}_{\text{worst},p-k}})$. ■

V. Reducing Search Time Using Satisfiability Modulo Theory Solving

Algorithm 2 exhaustively explores all combinations of $p-k$ sensors until a set $\mathbf{s}^*$ satisfying $d_{\text{attack},\mathbf{s}^*}(t_1) = 0$ is found. In this section, we explore the idea of using sophisticated search techniques in order to harness the underlying combinatorial aspect of the secure state estimation problem. In particular, we extend previous work by the authors and co-workers on using Satisfiability Modulo Theory (SMT)-like solvers [6], developed for the noiseless case, in order to improve the search time while preserving optimality of the solution.

The driving concept behind SMT solvers can be summarized as follows. First, the search space of all sensor subsets with cardinality $p-k$, is encoded using Boolean variables (the number of Boolean variables increases linearly with the number of sensors), and a Boolean search engine (*e.g.*, SAT solver) is used in order to traverse the search space. Whenever the SAT solver suggests one possible solution in the search space, a higher level solver (typically referred to as the Theory-solver) is used to check the correctness of that particular solution. Finally, in order to prevent the SAT solver from enumerating all possible solutions in the search space, the Theory-solver generates counter examples (certificates), explaining why a particular solution is not valid. Each certificate is used by the SAT solver in order to prune the search space and hence enhance the performance of the overall algorithm. This methodology of "counter-example guided search" effectively breaks the secure state estimation problem into two simpler tasks over the Boolean and Reals domain. Further details about this technique are described below.

A. Overall Architecture

We start by introducing a Boolean indicator variable $b = (b_1, \ldots, b_p) \in \mathbb{B}^p$ where the assignment $b_i = 1$ hypothesizes that the ith sensor is under attack while the assignment $b_i = 0$ hypothesizes that the ith sensor is attack-free. Using this indicator variable, b, we start by asking the (pseudo-)Boolean SAT solver to assign values to b in order to satisfy the following formula:

$$\phi(0) ::= \sum_{i=1}^{p} b_i \leq k, \tag{21}$$

which ensures that at most k sensors are going to be hypothesized as being under attack (the addition in (21) is over Reals).

In the next step, this hypothesized assignment is then checked by the theory solver. This is done by running the Attack-Detect algorithm (Algorithm 1) using only the set of hypothesized attack-free sensors $\mathbf{s}(b) = \{1,\ldots,p\} - \text{supp}(b)$. If the Attack-Detect algorithm returns $\hat{d}_{\text{attack},\mathbf{s}(b)} = 0$ then our solver approves this hypothesis and the algorithm terminates. Otherwise, an UNSAT certificate (also known as a counter-example) is generated explaining why this assignment of b is not valid (*i.e.*, a conflict). A trivial UNSAT certificate that can always be generated takes the following form (in iteration j):

$$\phi_{\text{cert}}(j) ::= \sum_{i\in\mathbf{s}(b)} b_i \geq 1, \tag{22}$$

which ensures that the current assignment of the variable b is excluded. Once this UNSAT certificate is generated, the (pseudo-) Boolean SAT solver is then invoked again in the next iteration with the following constraints:

$$\phi(j+1) ::= \phi(j) \wedge \phi_{\text{cert}}(j),$$

until one assignment of the variable b passes the attack detection test. This procedure is summarized in Algorithm 3.

Algorithm 3: SMT-Based Search.

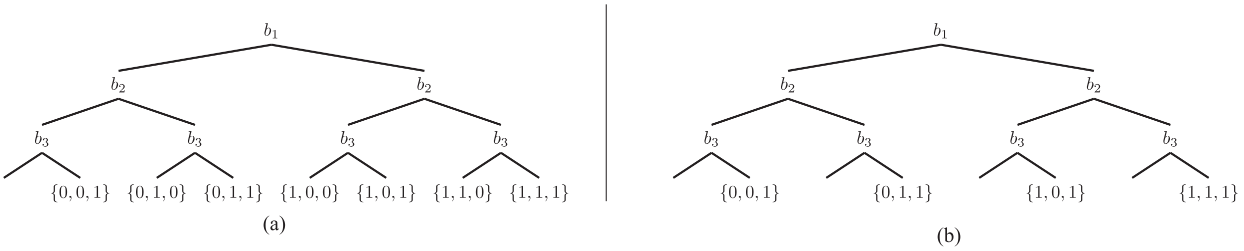
```
status := UNSAT;
$\phi_B := \sum_{i\in\{1,\ldots,p\}} b_i \leq k$;
**while** status == UNSAT **do**
    $b :=$ SAT-Solve($\phi_B$);
    $\mathbf{s}(b) := \{1,2,\ldots,p\} - \text{supp}(b)$;
    $(\hat{d}_{\text{attack},\mathbf{s}(b)}, \{\hat{\mathbf{x}}_{\mathbf{s}(b)}(t)\}_{t\in G})$
        := Attack-Detect($\mathbf{s}(b), t_1$);
    **if** $\hat{d}_{\text{attack},\mathbf{s}(b)} == 1$ **then**
        $\phi_{\text{cert}}$
            := Generate-Certificate($\mathbf{s}(b), \{\hat{\mathbf{x}}_{\mathbf{s}(b)}(t)\}_{t\in G}$);
        $\phi_B := \phi_B \wedge \phi_{\text{cert}}$;
    **end if**
**end while**
$\mathbf{s}^* = \mathbf{s}(b)$;
**return** $\{\hat{\mathbf{x}}_{\mathbf{s}^*}(t)\}_{t\in G}$;
```

B. Conflicting Certificates

The generated UNSAT certificates heavily affect the overall execution time. Smaller UNSAT certificates prune the search space faster. For simplicity, consider the example shown in Fig. 1 where the vector b has only three elements. On one hand, an UNSAT certificate that has the form $\phi_{\text{cert}} = b_1 + b_2 + b_3 \geq 1$ leads to pruning only one sample in the search space. On the other hand, a smaller UNSAT certificate that has the form $\phi_{\text{cert}} = b_1 \geq 1$ eliminates four samples in the search space which is indeed a higher reduction, and hence leads to better execution time.

To generate a compact (*i.e.*, smaller) Boolean constraint that explains a conflict, we aim to find a small set of sensors that cannot all be attack-free. To do so, we start by removing one sensor from the set $\mathbf{s}(b)$ and run the Attack-Detect algorithm on the reduced set $\mathbf{s}'(b)$ to obtain $\hat{d}_{\text{attack},\mathbf{s}'(b)}$. If $\hat{d}_{\text{attack},\mathbf{s}'(b)}$ still equals one (which indicates that set $\mathbf{s}'(b)$ still contains a conflicting set of sensors), we generate the more compact certificate:

$$\phi_{\text{cert}}(j) ::= \sum_{i\in\mathbf{s}'(b)} b_i \geq 1. \tag{23}$$



(a)

(b)

Fig. 1. Pictorial example illustrating the effect of generating smaller conflicting certificates. (a) A tree showing all the combinations of three Boolean indicator variables b_1, b_2, b_3 when a conflicting certificate of the form $\phi_{\text{cert}} := b_1 + b_2 + b_3 \geq 1$ is generated. The missing combination $\{0,0,0\}$ is the only one that is eliminated as a result of this certificate. (b) A tree showing all the combinations of three Boolean indicator variables b_1, b_2, b_3 when a conflicting certificate of the form $\phi_{\text{cert}} := b_3 \geq 1$ is generated. The missing four combinations $\{0,0,0\}$, $\{0,1,0\}$, $\{1,0,0\}$, $\{1,1,0\}$ are eliminated as a result of this certificate.

```
Algorithm 4: GENERATE-CERTIFICATE $(\mathbf{s}, \{\hat{\mathbf{x}}_{\mathbf{s}}(t)\}_{t\in G})$.
Compute the residues for $i \in \mathbf{s}$
$\mathbf{r}_i(t) := \mathbf{y}_i(t) - \mathcal{O}_i\hat{\mathbf{x}}_{\mathbf{s}}(t), \forall t \in G = \{t_1, \ldots, t_1+N-1\}$
$\mu_i(t_1) := \left|tr\left(\mathbb{E}_{N,t_1}\left(\mathbf{r}_i\mathbf{r}_i^T\right) - \mathcal{O}_i\mathbf{P}^\star_{\mathbf{s}}\mathcal{O}_i^T - \mathbf{M}_i\right) - \eta n\right|$;
Normalize the residues
$\mu_i(t_1) := \mu_i(t_1)/\lambda_{\max}\left(\mathcal{O}_i^T\mathcal{O}_i\right)$,
$\boldsymbol{\mu}(t_1) := \{\mu_i(t_1)\}_{i\in\mathbf{s}}$ ;
Sort the residues in ascending order
    $\boldsymbol{\mu}\_sorted(t_1)$ := sortAscendingly($\boldsymbol{\mu}(t_1)$);
Choose sensor indices of $p-2k+1$ smallest residues
    $\boldsymbol{\mu}\_min\_r$ := Index ($\mu\_sorted[1 : p-2k+1]$) ;
Search linearly for the UNSAT certificate
status = UNSAT; counter = 1; $\phi_{\text{conf-cert}} = 1$; $\mathbf{s}' = \mathbf{s}$
while status == UNSAT do
    $\mathbf{s}' := \mathbf{s}' \setminus \boldsymbol{\mu}\_\min\_r$[counter];
    $(\hat{d}_{\text{attack},\mathbf{s}'}, \{\hat{\mathbf{x}}_{\mathbf{s}'}(t)\}_{t\in G})$ := ATTACK-DETECT$(\mathbf{s}', t_1)$;
    if $\hat{d}_{\text{attack},\mathbf{s}'} == 1$ then
        $\phi_{\text{conf-cert}} := \phi_{\text{conf-cert}} \wedge \sum_{i\in\mathbf{s}'} b_i \geq 1$;
        counter := counter + 1;
    else
        status := SAT;
    end if
end while
return $\phi_{\text{conf-cert}}$
```

We continue removing sensors one by one until we cannot find any more conflicting sensor sets. Indeed, the order in which the sensors are removed is going to affect the overall execution time. In Algorithm 4 we implement a heuristic (for choosing this order) which is inspired by the strategy we adopted in the noiseless case [6].

Note that the reduced sets $\mathbf{s}'(b)$ are used only to generate the UNSAT certificates. Hence, it is direct to show that Algorithm 3 still preserves the optimality of the state estimate as stated by the following result.

Theorem 4: Let the linear dynamical system defined in (2) be $2k$-sparse observable system. Consider a k-adversary satisfying Assumptions 1–5. Consider the state estimate $\hat{\mathbf{x}}_{\mathbf{s}^*}(t)$ computed by Algorithm 3. Then, for any $\epsilon > 0$ and $\delta > 0$, there exists a large enough N such that:

$$\mathbb{P}\left(tr\left(\mathbb{E}_{N,t_1}\left(\mathbf{e}_{\mathbf{s}^*}\mathbf{e}_{\mathbf{s}^*}^T\right)\right) \leq tr\left(\mathbf{P}^\star_{\mathbf{s}_{\text{worst},\,p-k}}\right) + \epsilon\right) \geq 1 - \delta. \tag{24}$$

Note that although, for the sake of brevity, we did not analyze analytically the worst case execution time (in terms on number of iterations) of Algorithm 3, we show numerical results in Section VI that support the claim that the proposed SMT-like solver works much better in practice compared to the exhaustive search procedure (Algorithm 2).

VI. Numerical Experiments

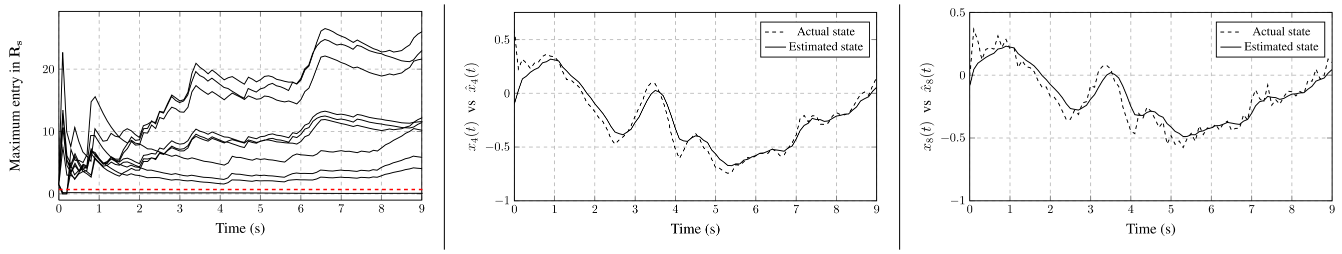
In this section, we report numerical results for Algorithms 2 and 3 as described by the experiments below.

A. *Experiment 1: Residue Test Performance in Algorithm 2*

In this experiment, we numerically check the performance of the residue test involved in Algorithm 2 while checking for effective attacks across sensor subsets. We generate a stable system randomly with $n = 20$ (state dimension) and $p = 5$ sensors. We select $k = 2$ sensors at random, and apply a random attack signal to the two sensors. We apply Algorithm 2 by running all the $\binom{5}{3} = 10$ Kalman filters (one for each distinct sensor subset of size 3) and do the residue test corresponding to each sensor subset. Fig. 2(a) shows the maximum entry in the residue test matrix $\mathbf{R}_{\mathbf{s}} = \mathbb{E}_{N,t_1}\left(\mathbf{r}_{\mathbf{s}}\mathbf{r}_{\mathbf{s}}^T\right) - \left(\mathcal{O}_{\mathbf{s}}\mathbf{P}^\star_{\mathbf{s}}\mathcal{O}_{\mathbf{s}}^T + \mathbf{M}_{\mathbf{s}}\right)$ for the 10 different Kalman filters. It is apparent from Fig. 2(a) that only one Kalman filter produces a state estimate that passes the residue test defined in Algorithm 1. This indeed corresponds to the set of attack-free sensors in the experiment.

B. *Experiment 2: Performance of SMT-Based Search*

In this experiment, we compare the sensor subset search time for the SMT-based approach (Algorithm 3) with that for the exhaustive search approach (Algorithm 2). For this experiment, we fix $n = 50$ (state dimension) and vary the number of sensors from $p = 3$ to $p = 15$. For each system, we pick one third of the sensors to be under attack, *i.e.*, $k = \lfloor p/3 \rfloor$. The attack signal is chosen as a linear function of the measurement noise. For each system, we run the bank of $\binom{p}{p-k}$ Kalman filters to generate the state estimates corresponding to all sensor subsets of size $p - k$. We then use both exhaustive search as well as the SMT-based search to find the sensor subset that satisfies the residue test in Algorithm 1. Fig. 3 shows the average time needed to perform the search across 50 runs of the same experiment. Fig. 3 suggests that the SMT-based search has an exponential improvement over exhaustive search as the number of sensors increases. In particular, for $p = 15$, the SMT-based search out-performs exhaustive search by an order of magnitude.



(a) (b)

Fig. 2. Figure showing results of Experiment 1: (a) the maximum entry in the residue test matrix $\mathbf{R_s} = \mathbb{E}_{N,t_1}\left(\mathbf{r_s}\mathbf{r_s}^T\right) - \left(\mathcal{O}_\mathbf{s}\mathbf{P}^\star_\mathbf{s}\mathcal{O}_\mathbf{s}^T + \mathbf{M_s}\right)$ for the 10 Kalman filters versus the threshold $\eta = 0.7$ (indicated by the dashed red line). As shown in the figure, there is only one subset of sensors which satisfies the threshold η, and this corresponds to the attack-free set of sensors, and (b) the estimated state trajectory (of state x_4 and x_8, *i.e.*, dimension 4 and 8 of $\mathbf{x}$) from the subset of sensors which satisfy the threshold versus the actual state trajectory.

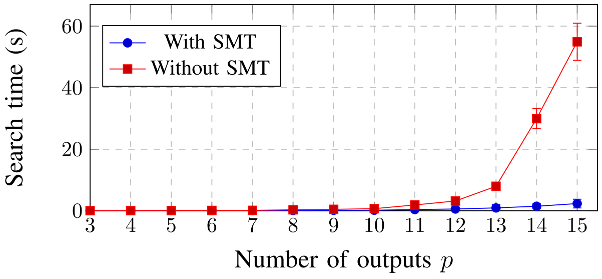


Fig. 3. Comparison of sensor subset search times for exhaustive search and SMT-based search.

VII. Sparse Observability: Coding Theoretic View

In this section, we revisit the sparse observability condition against a k-adversary and give a coding theoretic interpretation for the same. We first describe our interpretation for a linear system, and then discuss how it can be generalized for nonlinear systems.

Consider the linear dynamical system in (2) without the process and sensor noise (*i.e.*, $\mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t)$, $\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{a}(t)$). If the system's initial state is $\mathbf{x}(0) \in \mathbb{R}^n$ and the system is θ-sparse observable, then clearly in the absence of sensor attacks, by observing the outputs from any $p - \theta$ sensors for n time instants ($t = 0, 1, \ldots, n-1$) we can exactly recover $\mathbf{x}(0)$ and hence, *exactly* estimate the state of the plant. A coding theoretic view of this can be given as follows. Consider the outputs from sensor $d \in \{1, 2, \ldots, p\}$ for n time instants as a symbol $\mathcal{Y}_d \in \mathbb{R}^n$. Thus, in the (symbol) observation vector $\mathcal{Y} = \begin{bmatrix}\mathcal{Y}_1 & \mathcal{Y}_2 \ldots \mathcal{Y}_p\end{bmatrix}$, due to θ-sparse observability, any $p - \theta$ symbols are sufficient (in the absence of attacks) to recover the initial state $\mathbf{x}(0)$. Now, let us consider the case of a k-adversary which can arbitrarily corrupt any k sensors. In the coding theoretic view, this corresponds to arbitrarily corrupting any k (out of p) symbols in the observation vector. Intuitively, based on the relationship between error correcting codes and the Hamming distance between codewords in classical coding theory [13], one can expect the recovery of the initial state despite such corruptions to depend on the (symbol) Hamming distance between the observation vectors corresponding to two distinct initial states (say $\mathbf{x}^{(1)}(0)$ and $\mathbf{x}^{(2)}(0)$ with $\mathbf{x}^{(1)}(0) \neq \mathbf{x}^{(2)}(0)$). In this context, the following lemma relates θ-sparse observability to the minimum Hamming distance between observation vectors in the absence of attacks.

Lemma 2: For a θ-sparse observable system, the minimum (symbol) Hamming distance between observation vectors corresponding to distinct initial states is $\theta + 1$.

Proof: Consider a system with p sensors, and observation vectors $\mathcal{Y}^{(1)}$ and $\mathcal{Y}^{(2)}$ corresponding to distinct initial states $\mathbf{x}^{(1)}(0)$ and $\mathbf{x}^{(2)}(0)$. Due to θ-sparse observability, at most $p - \theta - 1$ symbols in $\mathcal{Y}^{(1)}$ and $\mathcal{Y}^{(2)}$ can be identical; if any $p - \theta$ of the symbols are identical, this would imply $\mathbf{x}^{(1)}(0) = \mathbf{x}^{(2)}(0)$. Hence, the (symbol) Hamming distance between the observation vectors $\mathcal{Y}^{(1)}$ and $\mathcal{Y}^{(2)}$ (corresponding to $\mathbf{x}^{(1)}(0)$ and $\mathbf{x}^{(2)}(0)$) is at least $p - (p - \theta - 1) = \theta + 1$ symbols. Also, there exists a pair of initial states $\left(\mathbf{x}^{(1)}(0), \mathbf{x}^{(2)}(0)\right)$, such that the corresponding observation vectors $\mathcal{Y}^{(1)}$ and $\mathcal{Y}^{(2)}$ are identical in exactly $p - \theta - 1$ symbols[3] and differ in the rest $\theta + 1$ symbols. Hence, the minimum (symbol) Hamming distance between the observation vectors is $\theta + 1$. ■

For a θ-sparse observable system, since the minimum Hamming distance between the observation vectors corresponding to distinct initial states is $\theta + 1$, we can:

1) correct up to $k < \frac{\theta+1}{2}$ sensor corruptions,
2) detect up to $k \leq \theta$ sensor corruptions.

Note that (1) above is equivalent to $2k \leq \theta$ (sparse observability condition for secure state estimation [4]). It should be noted that a k-adversary can attack *any* set of k (out of p) sensors, and the condition $k < \frac{\theta+1}{2}$ is both necessary and sufficient for exact state estimation despite such attacks. When $k \geq \frac{\theta+1}{2}$, it is straightforward to show a scenario where the observation vector (after attacks) can be explained by multiple initial states, and hence exact state estimation is not possible. The following example illustrates such an attack scenario.

Example 2: Consider a θ-sparse observable system with $\theta = 3$, number of sensors $p = 5$, and a k-adversary with $k = 2$. Clearly, the condition $k < \frac{\theta+1}{2}$ is not satisfied in this example. Let $\mathbf{x}^{(1)}(0)$ and $\mathbf{x}^{(2)}(0)$ be distinct initial states, such that the corresponding observation vectors $\mathcal{Y}^{(1)}$ and $\mathcal{Y}^{(2)}$ have (minimum) Hamming distance $\theta + 1 = 4$ symbols. Fig. 4 depicts the observation vectors $\mathcal{Y}^{(1)}$ and $\mathcal{Y}^{(2)}$, and for the sake of this example, we assume that the observation vectors have the same first symbol (i.e., $\mathcal{Y}_1^{(1)} = \mathcal{Y}_1^{(2)} = \mathcal{Y}_1$) and differ in the rest

[3] If there is no such pair of initial states, the initial state can be recovered by observing any $p - \theta - 1$ sensors. By definition, in a θ-sparse observable system, θ is the largest positive integer, such that the initial state can be recovered by observing any $p - \theta$ sensors.

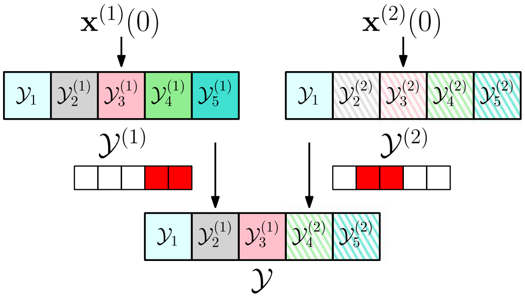

$\mathbf{x}^{(1)}(0)$
$\mathbf{x}^{(2)}(0)$
$\mathcal{Y}_1$ $\mathcal{Y}_2^{(1)}$ $\mathcal{Y}_3^{(1)}$ $\mathcal{Y}_4^{(1)}$ $\mathcal{Y}_5^{(1)}$
$\mathcal{Y}_1$ $\mathcal{Y}_2^{(2)}$ $\mathcal{Y}_3^{(2)}$ $\mathcal{Y}_4^{(2)}$ $\mathcal{Y}_5^{(2)}$
$\mathcal{Y}^{(1)}$
$\mathcal{Y}^{(2)}$
$\mathcal{Y}_1$ $\mathcal{Y}_2^{(1)}$ $\mathcal{Y}_3^{(1)}$ $\mathcal{Y}_4^{(2)}$ $\mathcal{Y}_5^{(2)}$
$\mathcal{Y}$

Fig. 4. Example with $\theta = 3$, $p = 5$ and $k = 2$. For distinct initial states $\mathbf{x}^{(1)}(0)$ and $\mathbf{x}^{(2)}(0)$, the corresponding observation vectors are $\mathcal{Y}^{(1)}$ and $\mathcal{Y}^{(2)}$. Given (attacked) observation vector $\mathcal{Y} = \left[\mathcal{Y}_1 \;\; \mathcal{Y}_2^{(1)} \;\; \mathcal{Y}_3^{(1)} \;\; \mathcal{Y}_4^{(2)} \;\; \mathcal{Y}_5^{(2)}\right]$, there are two possibilities for the initial state: (a) $\mathbf{x}^{(1)}(0)$ with attacks on sensors 4 and 5, or (b) $\mathbf{x}^{(2)}(0)$ with attacks on sensors 2 and 3.

4 symbols (hence, a Hamming distance of 4). Now, as shown in Fig. 4, suppose the observation vector after attacks was $\mathcal{Y} = \left[\mathcal{Y}_1 \;\; \mathcal{Y}_2^{(1)} \;\; \mathcal{Y}_3^{(1)} \;\; \mathcal{Y}_4^{(2)} \;\; \mathcal{Y}_5^{(2)}\right]$. Clearly, there are two possible explanations for this (attacked) observation vector: (a) the initial state was $\mathbf{x}^{(1)}(0)$ and sensors 4 and 5 were attacked, or (b) the initial state was $\mathbf{x}^{(2)}(0)$ and sensors 2 and 3 were attacked. Since there are two possibilities, we cannot estimate the initial state exactly given the attacked observation vector. This example can be easily generalized to show the necessity of the condition $k < \frac{\theta+1}{2}$.

For (noiseless) non-linear systems, by analogously defining θ-sparse observability, the same coding theoretic interpretation holds. This leads to the necessary and sufficient conditions for attack detection and secure state estimation in any noiseless dynamical system with sensor attacks.

Appendix

A. *Proof Details for Theorem 2*

1) Proof of (16) Using LLN:

$$tr\left(\mathbf{M}_{\mathbf{s}_g}\right) - tr\left(\mathbb{E}_{N,t_1}\left(\mathbf{z}_{\mathbf{s}_g}\mathbf{z}_{\mathbf{s}_g}^T\right)\right) \overset{(a)}{=} \sum_{l=0}^{n-1}\frac{1}{n}\left(tr\left(\mathbf{M}_{\mathbf{s}_g}\right) - \frac{1}{N_B}\sum_{t\in G_l} tr\left(\mathbf{z}_{\mathbf{s}_g}(t)\mathbf{z}_{\mathbf{s}_g}^T(t)\right)\right) \overset{(b)}{\leq} \epsilon_1,$$

where (a) follows from partitioning time window G (of size N) into n groups $G_0, G_1, \ldots G_{n-1}$ (each of size N_B) such that $G_l = \{t|\left((t-t_1) \;\; mod\; n\right) = l\}$, and (b) follows w.h.p. from LLN (for different time indices in G_l, $tr\left(\mathbf{z}_{\mathbf{s}_g}(t)\mathbf{z}_{\mathbf{s}_g}^T(t)\right)$ corresponds to i.i.d. realizations of the same random variable).

2) Cross Term Analysis and proof of (17): The cross term $2\mathbb{E}_{N,t_1}\left(\mathbf{z}_{\mathbf{s}_g}^T\mathcal{O}_{\mathbf{s}_g}\mathbf{e}_\mathbf{s}\right)$ can be written down as a sum of n terms as shown below:

$$2\mathbb{E}_{N,t_1}\left(\mathbf{z}_{\mathbf{s}_g}^T\mathcal{O}_{\mathbf{s}_g}\mathbf{e}_\mathbf{s}\right) \overset{(a)}{=} \frac{2}{n}\sum_{l=0}^{n-1}\left(\frac{1}{N_B}\sum_{t\in G_l}\mathbf{z}_{\mathbf{s}_g}^T(t)\mathcal{O}_{\mathbf{s}_g}\mathbf{e}_\mathbf{s}(t)\right) = \frac{2}{n}\sum_{l=0}^{n-1}\zeta_l, \quad (25)$$

where (a) follows from partitioning time window G (of size N) into n groups $G_0, G_1, \ldots G_{n-1}$ (each of size N_B) such that $G_l = \{t|\left((t-t_1) \;\; mod\; n\right) = l\}$. Now, we will show that each ζ_l has zero mean and vanishingly small variance for large enough N. The mean analysis can be done as shown below:

$$\mathbb{E}\left(\zeta_l\right) \overset{(a)}{=} \frac{1}{N_B}\sum_{t\in G_l}\mathbb{E}\left(\mathbf{z}_{\mathbf{s}_g}^T(t)\right)\mathbb{E}\left(\mathcal{O}_{\mathbf{s}_g}\mathbf{e}_\mathbf{s}(t)\right) = 0, \quad (26)$$

where (a) follows from the independence of $\mathbf{e}_\mathbf{s}(t)$ from $\mathbf{z}_{\mathbf{s}_g}^T(t)$ (due to Assumptions 4 and 5). This implies that the cross term $2\mathbb{E}_{N,t_1}\left(\mathbf{z}_{\mathbf{s}_g}^T\mathcal{O}_{\mathbf{s}_g}\mathbf{e}_\mathbf{s}\right)$ has zero mean. From (26) and (16),

$$\begin{aligned}2\epsilon_1 &\geq \mathbb{E}\left(\mathbb{E}_{N,t_1}\left(tr\left(\mathcal{O}_{\mathbf{s}_g}\left(\mathbf{e}_\mathbf{s}\mathbf{e}_\mathbf{s}^T - \mathbf{P}_\mathbf{s}^\star\right)\mathcal{O}_{\mathbf{s}_g}^T\right)\right)\right)\\ &= \mathbb{E}\left(\mathbb{E}_{N,t_1}\left(tr\left(\left(\mathbf{e}_\mathbf{s}\mathbf{e}_\mathbf{s}^T - \mathbf{P}_\mathbf{s}^\star\right)\mathcal{O}_{\mathbf{s}_g}^T\mathcal{O}_{\mathbf{s}_g}\right)\right)\right)\\ &\overset{(a)}{\geq} \lambda_{\min}\left(\mathcal{O}_{\mathbf{s}_g}^T\mathcal{O}_{\mathbf{s}_g}\right)\mathbb{E}\left(\mathbb{E}_{N,t_1}\left(tr\left(\mathbf{e}_\mathbf{s}\mathbf{e}_\mathbf{s}^T - \mathbf{P}_\mathbf{s}^\star\right)\right)\right), \quad (27)\end{aligned}$$

where (a) follows from Lemma 3 (discussed in Appendix B).

Using (27), we can show that for any $\epsilon_2 > 0$, there exists a large enough N_B such that (see [15] for details):

$$\mathbb{E}\left(\zeta_l^2\right) = \mathbb{E}\left(\left(\frac{1}{N_B}\sum_{t\in G_l}\mathbf{e}_\mathbf{s}^T(t)\mathcal{O}_{\mathbf{s}_g}^T\mathbf{z}_{\mathbf{s}_g}(t)\right)^2\right) \leq \epsilon_2. \quad (28)$$

Clearly ζ_l has vanishingly small variance as $N_B \to \infty$. As a consequence, the variance of the cross term $2\mathbb{E}_{N,t_1}(\mathbf{z}_{\mathbf{s}_g}^T\mathcal{O}_{\mathbf{s}_g}\mathbf{e}_\mathbf{s}) = \frac{2}{n}\sum_{l=0}^{n-1}\zeta_l$ is also vanishingly small for $N_B \to \infty$ (follows from the Cauchy-Schwarz inequality). Since the cross term $2\mathbb{E}_{N,t_1}(\mathbf{z}_{\mathbf{s}_g}^T\mathcal{O}_{\mathbf{s}_g}\mathbf{e}_\mathbf{s})$ has zero mean and vanishingly small variance, by the Chebyshev inequality, $|2\mathbb{E}_{N,t_1}(\mathbf{z}_{\mathbf{s}_g}^T\mathcal{O}_{\mathbf{s}_g}\mathbf{e}_\mathbf{s})| \leq \epsilon_1$ holds w.h.p., and this completes the proof of (17).

B. *Bounds on the Trace of Product of Symmetric Matrices*

Lemma 3: If $\mathbf{A}$ and $\mathbf{B}$ are two symmetric matrices in $\mathbb{R}^{n\times n}$, and $\mathbf{B}$ is positive semi-definite:

$$\lambda_{\min}\left(\mathbf{A}\right) tr\left(\mathbf{B}\right) \leq tr\left(\mathbf{AB}\right) \leq \lambda_{\max}\left(\mathbf{A}\right) tr\left(\mathbf{B}\right). \quad (29)$$

References

[1] S. Mishra, Y. Shoukry, N. Karamchandani, S. Diggavi, and P. Tabuada, “Secure state estimation: Optimal guarantees against sensor attacks in the presence of noise,” in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2015.

[2] F. Pasqualetti, F. Dorfler, and F. Bullo, “Control-theoretic methods for cyberphysical security: Geometric principles for optimal cross-layer resilient control systems,” *IEEE Control Syst.*, vol. 35, no. 1, pp. 110–127, Feb. 2015.

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

[4] Y. Shoukry and P. Tabuada, "Event-triggered state observers for sparse sensor noise/attacks," *arXiv pre-print*, Sep. 2013.. [Online]. Available: http://arxiv.org/abs/1309.3511.

[5] F. Pasqualetti, F. Dörfler, and F. Bullo, "A divide-and-conquer approach to distributed attack identification," in *Proc. IEEE Conf. Dec. Control*, Dec. 2015.

[6] Y. Shoukry, P. Nuzzo, A. Puggelli, A. L. Sangiovanni-Vincentelli, S. A. Seshia, and P. Tabuada, "Secure state estimation for cyber physical systems under sensor attacks: A satisfiability modulo theory approach," *arXiv pre-print*, Dec. 2014.

[7] M. S. Chong, M. Wakaiki, and J. P. Hespanha, "Observability of linear systems under adversarial attacks," in *Proc. Amer. Control Conf.*, 2015.

[8] M. Pajic, J. Weimer, N. Bezzo, P. Tabuada, O. Sokolsky, I. Lee, and G. Pappas, "Robustness of attack-resilient state estimators," in *Proc. ACM/IEEE Int. Conf. Cyber-Phys. Syst.*, 2014.

[9] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *Proc. Allerton Conf. Commun., Control, Comput.*, 2009.

[10] C.-Z. Bai and V. Gupta, "On kalman filtering in the presence of a compromised sensor: Fundamental performance bounds," in *Proc. Amer. Control Conf.*, 2014.

[11] J. Mattingley and S. Boyd, "Real-time convex optimization in signal processing," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 50–61, May 2010.

[12] S. Farahmand, G. B. Giannakis, and D. Angelosante, "Doubly robust smoothing of dynamical processes via outlier sparsity constraints," *IEEE Trans. Signal Process.*, vol. 59, no. 10, pp. 4529–4543, Oct. 2011.

[13] R. Blahut, *Algebraic Codes for Data Transmission*. Cambridge University Press, 2003.

[14] T. Kailath, A. Sayed, and B. Hassibi, *Linear Estimation*. Prentice-Hall, 2000.

[15] S. Mishra, Y. Shoukry, N. Karamchandani, S. Diggavi, and P. Tabuada, "Secure state estimation against sensor attacks in the presence of noise," *arXiv pre-print*, Oct. 2015. [Online]. Available: http://arxiv.org/abs/1510.02462

[16] A. Willsky, "A survey of design methods for failure detection in dynamic systems," *Automatica*, vol. 12, no. 6, pp. 601–611, Nov. 1976.

Shaunak Mishra received the B.Tech. degree in electronics and electrical communication engineering from the Indian Institute of Technology, Kharagpur, India, in 2010, the M.S. degree in electrical engineering from the University of California, Los Angeles, CA, USA, in 2011, and is currently pursuing the Ph.D. degree in electrical engineering at University of California, Los Angeles, CA, USA.

His research interests include statistics and information theory with applications in security and machine learning.

Dr. Mishra is a recipient of the Henry Samueli fellowship at UCLA.

Yasser Shoukry received the B.Sc. and M.Sc. degrees (Hons.) in computer and systems engineering from Ain Shams University, Cairo, Egypt, in 2007 and 2010, respectively, and the Ph.D. degree in electrical engineering from the University of California at Los Angeles (UCLA), Los Angeles, CA, USA, in 2015, where he was affiliated with both the Cyber-Physical Systems Lab as well as the Networked and Embedded Systems Lab.

Currently, he is a Postdoctoral Scholar in the Department of Electrical Engineering and Computer Sciences of the University of California, Berkeley, CA, USA, and the Department of Electrical Engineering, UCLA. Before joining UCLA, he spent four years as an R&D Engineer in the industry of automotive embedded systems. His research interests include the design and implementation of secure and privacy aware cyberphysical systems by drawing on tools from control theory, optimization theory, embedded systems, and formal methods.

Dr. Shoukry is the recipient of the Best Paper Award from the International Conference on Cyber Physical Systems (ICCPS) in 2016, as well as the the UCLA EE Preliminary Exam Fellowship, the UCLA Chancellors prize, the UCLA EE Graduate Division Fellowship, and Distinguished Dissertation Award in 2011, 2012, and 2016.

Nikhil Karamchandani received the M.S. and Ph.D. degrees in electrical and computer engineering from the University of California at San Diego, San Diego, CA, USA, in 2007 and 2011, respectively.

From 2011 to 2014, he was a Postdoctoral Scholar at the University of California at Los Angeles and at the Information Theory and Applications (ITA) Center at the University of California at San Diego. Currently, he is an Assistant Professor in the Department of Electrical Engineering at the Indian Institute of Technology Bombay, Mumbai, India. His research interests are in networks, communications, and information theory.

Dr. Karamchandani received the California Institute for Telecommunications and Information Technology (CalIT2) fellowship in 2005 and the INSPIRE Faculty Fellowship in 2015.

Suhas N. Diggavi (F'13) received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Delhi, India, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 1998.

After completing his Ph.D. degree, he was a Principal Member Technical Staff with the Information Sciences Center, AT&T Shannon Laboratories, Florham Park, NJ, USA. After that, he was on the faculty of the School of Computer and Communication Sciences, EPFL, where he directed the Laboratory for Information and Communication Systems (LICOS). Currently, he is a Professor in the Department of Electrical Engineering at the University of California, Los Angeles, CA, USA, where he directs the Information Theory and Systems Laboratory. His research interests include wireless network information theory, wireless networking systems, as well as network data compression and network algorithms.

Dr. Diggavi has received several recognitions for his research including the 2013 IEEE Information Theory Society & Communications Society Joint Paper Award, the 2013 ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc) best paper award, and the 2006 IEEE Donald Fink prize paper award. He is currently a Distinguished Lecturer and serves on Board of Governors for the IEEE Information Theory Society. He has been an associate editor for IEEE Transactions on Information Theory, ACM/IEEE Transactions on Networking, IEEE Communication Letters, a guest editor for IEEE Selected Topics in Signal Processing and of the program committees of several IEEE conferences. He has also helped organize IEEE conferences including serving as the Technical Program Co-Chair for 2012 IEEE Information Theory Workshop (ITW) and the Technical Program Co-Chair for the 2015 IEEE International Symposium on Information Theory (ISIT). He has eight issued patents.

Paulo Tabuada was born in Lisbon, Portugal. He received the "Licenciatura" degree in aerospace engineering from Instituto Superior Tecnico, Lisbon, Portugal, in 1998 and the Ph.D. degree in electrical and computer engineering from the Institute for Systems and Robotics, in 2002, a private research institute associated with Instituto Superior Tecnico.

He was a Postdoctoral Researcher at the University of Pennsylvania, University Park, PA, USA, from 2002 to 2003. After spending three years at the University of Notre Dame, Notre Dame, IN, USA, as an Assistant Professor, he joined the Electrical Engineering Department at the University of California, Los Angeles, CA, USA, where he established and directs the Cyber-Physical Systems Laboratory.

Prof. Tabuada's contributions to cyberphysical systems have been recognized by multiple awards, including the National Science Foundation CAREER award in 2005, the Donald P. Eckman award in 2009, the George S. Axelby award in 2011, and the Antonio Ruberti Prize in 2015. In 2009, he co-chaired the International Conference Hybrid Systems: Computation and Control (HSCC'09) and joined its steering committee in 2015. In 2012, he was Program Co-Chair for the 3rd IFAC Workshop on Distributed Estimation and Control in Networked Systems (NecSys'12), and in 2015, he was Program Co-Chair for the IFAC Conference on Analysis and Design of Hybrid Systems. He also served on the editorial board of the IEEE Embedded Systems Letters and the IEEE Transactions on Automatic Control. His latest book on verification and control of hybrid systems has been published (Springer, 2009).