Secure State Estimation for Cyber-Physical Systems Under Sensor Attacks: A Satisfiability Modulo Theory Approach

Yasser Shoukry, Pierluigi Nuzzo, *Member, IEEE*, Alberto Puggelli, *Student Member, IEEE*, Alberto L. Sangiovanni-Vincentelli, Sanjit A. Seshia, *Senior Member, IEEE*, and Paulo Tabuada

***Abstract*—Secure state estimation is the problem of estimating the state of a dynamical system from a set of noisy and adversarially corrupted measurements. Intrinsically a combinatorial problem, secure state estimation has been traditionally addressed either by brute force search, suffering from scalability issues, or via convex relaxations, using algorithms that can terminate in polynomial time but are not necessarily sound. In this paper, we present a novel algorithm that uses a satisfiability modulo theory approach to harness the complexity of secure state estimation. We leverage results from formal methods over real numbers to provide guarantees on the soundness and completeness of our algorithm. Moreover, we discuss its scalability properties, by providing upper bounds on the runtime performance. Numerical simulations support our arguments by showing an order of magnitude decrease in execution time with respect to alternative techniques. Finally, the effectiveness of the proposed algorithm is demonstrated by applying it to the problem of controlling an unmanned ground vehicle.**

***Index Terms*—Secure cyber-physical systems, secure state estimation, sensor attacks, satisfiability modulo theories.**

I. Introduction

THE detection and mitigation of attacks on cyber-physical systems (CPS) is a problem of increasing importance. The tight coupling between "cyber" components and "physical" processes often leads to systems where the increased sophistication comes at the expense of increased vulnerability and security weaknesses. An important scenario is posed by a malicious adversary that can arbitrarily corrupt the measurements of a subset of sensors in the system. These sensor-related attacks can be deployed by using either cyber or physical components as follows:

1) *Software.* Malicious software running on the processor executing the sensor processing routine can access the sensor information before it is processed by the controller itself. The Stuxnet malware is an infamous example of this category of attacks. It exploits vulnerabilities in the operating system running over supervisory control and data acquisition (SCADA) devices [1] and once it obtains enough operating system privileges, it can corrupt the sensor measurements collected via the attacked SCADA device.
2) *Network.* Modern control systems rely on a networked infrastructure to exchange sensor information. Therefore, an adversarial attacker can corrupt sensor measurements by manipulating the data packets exchanged between various components, as has been investigated, for instance, in smart grids [2].
3) *Sensors Spoofing.* By tampering with the sensor hardware or environment, an adversary can mislead the sensor about the value of the physical signal it is attempting to measure. As previously shown by some of the authors, it is possible to make drivers lose control of their cars by directly spoofing the velocity sensors of antilock braking systems in a noninvasive manner [3].

In all the scenarios above, because sensor measurements are used to generate control commands, corrupted measurements can lead to corrupted commands, thus critically affecting the physical process under control.

This paper addresses the problem of estimating the state of the underlying physical system from corrupted measurements, so that it can be used by the controller. We call this problem *secure state estimation*. We focus on linear dynamical systems and model the attack as a sparse vector added to the measurement vector. The entries corresponding to unattacked sensors are null while sensors under attack are corrupted by nonzero signals. We make no assumptions regarding the magnitude, statistical description, or temporal evolution of the attack vector.

Manuscript received October 20, 2015; revised May 9, 2016 and January 17, 2017; accepted February 7, 2017. Date of publication March 1, 2017; date of current version September 25, 2017. This work was supported in part by the National Science Foundation under Grant 1136174, in part by the DARPA under Agreement FA8750-12-2-0247, in part by TerraSwarm, one of six centers of STARnet, a Semiconductor Research Corporation program sponsored by MARCO and DARPA, and in part by the National Science Foundation project ExCAPE: Expeditions in Computer Augmented Program Engineering (Award 1138996). Recommended by Associate Editor H. Lin.

Y. Shoukry and P. Tabuada are with the Electrical Engineering Department, University of California, Los Angeles, Los Angeles, CA 90095 USA (e-mail: yshoukry@ucla.edu; tabuada@ucla.edu).

P. Nuzzo is with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90007 USA (e-mail: nuzzo@usc.edu).

A. Puggelli is with the Lion Semiconductor Inc., San Francisco, CA 94107 USA (e-mail: alberto@lionsemi.com).

A. L. Sangiovanni-Vincentelli and S. A. Seshia are with the Department of Electrical Engineering and Computer Sciences, University of California Berkeley, Berkeley, CA 94720 USA (e-mail: alberto@eecs.berkeley.edu; sseshia@eecs.berkeley.edu).

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/TAC.2017.2676679

While prior work has addressed the secure state estimation problem for the special cases of scalar systems [4], [5], or when the attack signal has a specific structure (e.g., in the case of replay attacks [6]), we focus instead on the general case, in which the system under attack is multidimensional, it is equipped with multiple sensors, and there are no assumptions on the time evolution of the attack signal. In this case, secure state estimation becomes a combinatorial problem [7]–[9]. We can then categorize the different contributions in the literature based on the techniques used to tackle the combinatorial aspects in it, namely, by brute force search [8]–[11] and by convex relaxations [7], [12], [13].

Pasqualetti *et al.* provide a suite of sound and complete algorithms to generate fault-monitor filters, which can be used to detect the existence of an attack [8]. However, if only an upper bound on the cardinality of the attacked sensors is available, the number of needed monitors is combinatorial in the size of the attacked sensors, which might hinder the scalability of the approach. A similar approach is employed by Yong *et al.* under the assumptions that both sensors and actuators are attacked and both the process dynamics and the sensors are affected by stochastic noise [14]. To avoid running a combinatorial set of parallel monitors, Chong *et al.* [9] show how all the monitors can be combined into a single multi-observer component. However, the number of the observer outputs is still combinatorial, and the proposed algorithm must exhaustively search over all of them to discover which sensors are under attack.

As an alternative approach, the secure state estimation problem can be formulated as a nonconvex l_0 minimization problem, and then relaxed into a convex l_1/l_r problem, which can be solved in polynomial time. This technique has been reported both in the case where sensors are ideal and not affected by noise [7] and in the noisy case [12]. However, a major drawback of such a relaxation step is the loss of correctness guarantees, as witnessed by some of the numerical results in this paper, in which the relaxed l_1/l_r formulation leads to incorrect estimates. Algorithms that can avoid the relaxation step, while running in polynomial time, have also been recently proposed [15], [16]; however, their correctness is only guaranteed under restrictive assumptions on the system structure.

Outside of the two categories above, an online learning mechanism based on approximate envelopes of collected data has been recently proposed for secure state estimation [17]. The envelopes are used to detect any abnormal behavior without assuming any knowledge of the dynamical system model. Robustification techniques were also used for state estimation against sparse sensor attacks, using either Kalman filters or principal component analysis [18], [19]. However, no formal guarantees on the correctness of these approaches are currently available. Our problem is also related to fault-tolerant state estimation and control. However, the literature on fault-tolerant control [20] typically relies on prior knowledge of the failure modes and the statistical properties or time evolution of failures, which is generally not available in the context of adversarial attacks.

In this paper, we resort to techniques from formal methods to develop a *sound and complete* algorithm that can *efficiently* handle the combinatorial complexity of the state estimation problem. We show that the state estimation problem can be cast as a satisfiability problem for a formula including logic and pseudo-Boolean constraints on Boolean variables as well as convex constraints on real variables. The Boolean variables model the presence (or absence) of an attack, while the convex constraints capture properties of the system state. We then show how this satisfiability problem can be efficiently solved using the *Satisfiability Modulo Theory* (SMT) paradigm [21], specifically adapted to convex constraint solving [22], to provide both the index of the attacked sensors and the state estimate. To improve the execution time of our decision procedure, we equip the convex constraint solver of our SMT-based algorithm with heuristics that can exploit the specific geometry of the state estimation problem while preserving soundness and completeness. Finally, we compare the performance of our approach against other algorithms via numerical experiments, and demonstrate its effectiveness on the problem of controlling an unmanned ground vehicle (UGV). Our technical contributions can be summarized as follows:

1) We provide a formalization of the secure state estimation problem as a satisfiability problem, which includes both Boolean constraints and convex constraints over real variables.
2) We develop IMHOTEP[1]-SMT, a novel SMT-solver that is shown to provide a sound and complete solution to the secure state estimation problem in the absence of measurement noise and model uncertainties.
3) We propose procedures to improve the execution time of the IMHOTEP-SMT solver along with upper bounds on the number of iterations required by the proposed algorithm.
4) We extend the analysis of IMHOTEP-SMT to the case when sensor measurements and model uncertainties are bounded. We show that, while the exact support of the attack may not always be identifiable, we can still guarantee and quantify the boundedness of the state estimation error.

We reported a preliminary version of these results in which only the special case of "perfect" model (i.e., the sensors are noiseless and there is no mismatch between the model and the actual system) was discussed, without providing the proofs of the formal guarantees [23]. A subsequent paper detailed the implementation of the proposed SMT-based solver [24]. In this paper, we discuss and prove in detail all the theoretical results used in our previous work [23], [24] and extend them to the case when uncertainties in the model as well as sensor noise are present.

The rest of this paper is organized as follows. Section II introduces the formal setup for the problem under consideration. The main contributions of this paper—the introduction of the SMT-based detector and the characterization of its soundness and completeness—are presented in Sections III and IV. Numerical comparisons and results are then reported in Section V. Finally, Section VI concludes the paper and discusses new research directions.

[1] IMHOTEP (pronounced as "emmo-tepp") was an ancient Egyptian polymath who is considered to be the earliest known architect, engineer, and physician in the early history. He is famous for the design of the oldest pyramid in Egypt, the Pyramid of Djoser (the Step Pyramid) at Saqqara, 2630–2611 BC.

II. Secure State Estimation Problem

We provide a mathematical formulation of the state estimation problem considered in this paper and discuss the conditions for the existence and uniqueness of its solution.

A. Notation

The symbols $\mathbb{N}, \mathbb{R}$, and $\mathbb{B}$ denote the sets of natural, real, and Boolean numbers, respectively. The symbols $\wedge$ and $\neg$ denote the logical AND and logical NOT operators, respectively. The support of a vector $x \in \mathbb{R}^n$, denoted by $\mathrm{supp}(x)$, is the set of indices of the nonzero elements of x. Similarly, the complement of the support of a vector x is denoted by $\overline{\mathrm{supp}(x)} = \{1, \ldots, n\} \setminus \mathrm{supp}(x)$. If S is a set, $|S|$ is the cardinality of S. We call a vector $x \in \mathbb{R}^n$ s-sparse, if x has at most s nonzero elements, i.e., if $|\mathrm{supp}(x)| \le s$.

Given p vectors of the same dimension $x_1, \ldots, x_p \in \mathbb{R}^n$, we call $x = (x_1, x_2, \ldots, x_p) \in \mathbb{R}^{pn}$ a block vector and each component x_i a block. To emphasize that a vector x is a block vector, we write it as an element of $\mathbb{R}^{pn}$, where the exponent pn is written as the juxtaposition of the number of blocks p and the size of individual blocks n, respectively. With some abuse of notation, for the block vector $x = (x_1, x_2, \ldots, x_p) \in \mathbb{R}^{pn}$, we denote by $\mathrm{supp}(x)$ the indices of the blocks on which $x \in \mathbb{R}^{pn}$ is supported. In other words, an index $i \in \{1, \ldots, p\}$ belongs to the set $\mathrm{supp}(x) \subseteq \{1, \ldots, p\}$ whenever the ith block x_i is nonzero, i.e.,

$$i \in \mathrm{supp}(x) \Leftrightarrow x_i \neq 0, \qquad i \in \{1, \ldots, p\}.$$

Similarly, a block matrix $M \in \mathbb{R}^{pn \times m}$ is defined as the vertical concatenation of the matrices $M_1, \ldots, M_p \in \mathbb{R}^{n \times m}$. In such case, a block is defined as the matrix $M_i \in \mathbb{R}^{n \times m}$, hence the matrix M can be written as $M = \begin{bmatrix} M_1^T & \ldots & M_p^T \end{bmatrix}^T$. Similarly to the notation used for vectors, the row dimension of the block matrix $M \in \mathbb{R}^{pn \times m}$ is written as the juxtaposition of the number of blocks p and the size of the individual blocks n.

For a vector $x \in \mathbb{R}^n$, we denote by $\|x\|_2$ the 2-norm of x and by $\|M\|_2$ the induced 2-norm of a matrix $M \in \mathbb{R}^{m \times n}$. We also denote by $M_i \in \mathbb{R}^{1 \times n}$ the ith row of M. For the set $\Gamma \subseteq \{1, \ldots, m\}$, we denote by $M_\Gamma \in \mathbb{R}^{|\Gamma| \times n}$ the matrix obtained from M by removing all the rows except those indexed by Γ. Then, $M_{\overline{\Gamma}} \in \mathbb{R}^{(m-|\Gamma|) \times n}$ is the matrix obtained from M by removing the rows indexed by the set Γ, $\overline{\Gamma}$ representing the complement of Γ. For example, if $m = 4$, and $\Gamma = \{1, 2\}$, we have

$$M_\Gamma = \begin{bmatrix} M_1 \\ M_2 \end{bmatrix}, \quad M_{\overline{\Gamma}} = \begin{bmatrix} M_3 \\ M_4 \end{bmatrix}.$$

By the same abuse of notation, for a block matrix $M \in \mathbb{R}^{pn \times m}$, we denote by $M_\Gamma \in \mathbb{R}^{|\Gamma| n \times m}$ the block matrix obtained by removing all blocks except those indexed by Γ. We define $M_{\overline{\Gamma}}$ similarly.

B. System and Attack Model

We consider a system under sensor attack of the form:

$$x^{(t+1)} = Ax^{(t)} + Bu^{(t)} + \mu^{(t)} \tag{1}$$

$$y^{(t)} = Cx^{(t)} + a^{(t)} + \psi^{(t)} \tag{2}$$

where $x^{(t)} \in \mathbb{R}^n$ is the system state at time $t \in \mathbb{N}$, $u^{(t)} \in \mathbb{R}^m$ is the system input, and $y^{(t)} \in \mathbb{R}^p$ is the observed output. The matrices A, B, and C represent the system dynamics and have appropriate dimensions. The attack vector $a^{(t)} \in \mathbb{R}^p$ is an s-sparse vector modeling how an attacker changed the sensor measurements at time t. If sensor $i \in \{1, \ldots, p\}$ is attacked then the ith element in $a^{(t)}$ is nonzero; otherwise the ith sensor is not attacked. Hence, s describes the number of attacked sensors. Note that we make no assumptions on the vector $a^{(t)}$ apart from being s-sparse. In particular, we do not assume bounds, statistical properties, nor restrictions on the time evolution of the elements in $a^{(t)}$. The value of s is also not assumed to be known, although we assume the knowledge of an upper bound $\overline{s}$ on the number of sensors that can be attacked. We, therefore, only assume that the attacker has access to a subset of sensors of cardinality $s \le \overline{s}$; whether a specific sensor in this subset is attacked or not may change with time. As shown in the following section, the maximum number of attacked sensors that can be detected is a characteristic of the system and depends on the pair (A, C). Finally, the vectors $\mu^{(t)}$ and $\psi^{(t)} \in \mathbb{R}^p$ represent, respectively, the process noise and the measurement noise, which are assumed to be uniformly bounded, i.e., there exist constants $\overline{\mu}$ and $\overline{\psi}$ such that the bounds $\left\|\mu^{(t)}\right\|_2 \le \overline{\mu}$ and $\left\|\psi^{(t)}\right\|_2 \le \overline{\psi}$ are satisfied for all time $t \in \mathbb{N}$.

C. Problem Formulation

To formulate the state estimation problem, we assume that the state is reconstructed from a set of τ measurements ($\tau \in \mathbb{N}$), where $\tau \le n$ is selected to guarantee that the system observability matrix, as defined below, has full rank. Therefore, we can arrange the outputs from the ith sensor at different time instants as follows:

$$\widetilde{Y}_i^{(t)} = \mathcal{O}_i x^{(t-\tau+1)} + E_i^{(t)} + F_i U^{(t)} + \Psi_i^{(t)}$$

where:

$$\widetilde{Y}_i^{(t)} = \begin{bmatrix} y_i^{(t-\tau+1)} \\ y_i^{(t-\tau+2)} \\ \vdots \\ y_i^{(t)} \end{bmatrix}, E_i^{(t)} = \begin{bmatrix} a_i^{(t-\tau+1)} \\ a_i^{(t-\tau+2)} \\ \vdots \\ a_i^{(t)} \end{bmatrix}, U^{(t)} = \begin{bmatrix} u^{(t-\tau+1)} \\ u^{(t-\tau+2)} \\ \vdots \\ u^{(t)} \end{bmatrix}$$

$$F_i = \begin{bmatrix} 0 & 0 & \ldots & 0 & 0 \\ C_i B & 0 & \ldots & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ C_i A^{\tau-2} B & C_i A^{\tau-3} B & \ldots & C_i B & 0 \end{bmatrix}, \mathcal{O}_i = \begin{bmatrix} C_i \\ C_i A \\ \vdots \\ C_i A^{\tau-1} \end{bmatrix}$$

$$\Psi_i^{(t)} = \begin{bmatrix} 0 & 0 & \ldots & 0 & 0 \\ C_i & 0 & \ldots & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ C_i A^{\tau-2} & C_i A^{\tau-3} & \ldots & C_i & 0 \end{bmatrix} \begin{bmatrix} \mu^{(t-\tau+1)} \\ \mu^{(t-\tau+2)} \\ \vdots \\ \mu^{(t)} \end{bmatrix} + \begin{bmatrix} \psi_i^{(t-\tau+1)} \\ \psi_i^{(t-\tau+2)} \\ \vdots \\ \psi_i^{(t)} \end{bmatrix}.$$

Since all the inputs in $U^{(t)}$ are known, we can further simplify the output equation as

$$Y_i^{(t)} = \mathcal{O}_i x^{(t-\tau+1)} + E_i^{(t)} + \Psi_i^{(t)} \tag{3}$$

where $Y_i^{(t)} = \widetilde{Y}_i^{(t)} - F_i U^{(t)}$. We also define the block vectors $Y^{(t)}, E^{(t)}, \Psi^{(t)} \in \mathbb{R}^{p\tau}$ and the block matrix $\mathcal{O} \in \mathbb{R}^{p\tau \times n}$ as

$$Y^{(t)} = \begin{bmatrix} Y_1^{(t)} \\ \vdots \\ Y_p^{(t)} \end{bmatrix}, E^{(t)} = \begin{bmatrix} E_1^{(t)} \\ \vdots \\ E_p^{(t)} \end{bmatrix}, \Psi^{(t)} = \begin{bmatrix} \Psi_1^{(t)} \\ \vdots \\ \Psi_p^{(t)} \end{bmatrix}, \mathcal{O} = \begin{bmatrix} \mathcal{O}_1 \\ \vdots \\ \mathcal{O}_p \end{bmatrix} \tag{4}$$

to denote, respectively, the vector of outputs, attacks, and observability matrices related with all sensors over the same time window of length τ. Note that, even if $\Psi_i^{(t)}$ represents both process and measurement noise, for the sake of simplicity, we will refer to $\Psi_i^{(t)}$ as measurement noise. It follows from the boundedness assumption on the process and measurement noise that there exist constants $\overline{\Psi}_1, \ldots, \overline{\Psi}_p$ such that the bound $\left\| \Psi_i^{(t)} \right\|_2 \leq \overline{\Psi}_i$ is always satisfied for all $t \in \mathbb{N}$ and for all sensors $i \in \{1, \ldots, p\}$. Finally, with some abuse of notation, for the set of indices $\mathcal{I} \subseteq \{1, \ldots, p\}$ we denote by $\overline{\Psi}_\mathcal{I}$ the noise bound of the sensors indexed by $\mathcal{I}$, i.e., $\left\| \Psi_\mathcal{I}^{(t)} \right\|_2 \leq \overline{\Psi}_\mathcal{I}$ with $\overline{\Psi}_\mathcal{I}^2 = \sum_{i \in \mathcal{I}} \overline{\Psi}_i^2$. With the same abuse of notation, we denote by $\overline{\Psi}$ the noise bound of all sensors, i.e., $\overline{\Psi}^2 = \sum_{i=1}^p \overline{\Psi}_i^2$.

D. Problem Statement

For each sensor, we define a binary indicator variable $b_i \in \mathbb{B}$ such that $b_i = 1$ when the ith sensor is under attack and $b_i = 0$ otherwise. Based on the formulation in Section II-C, our goal is to find $x^{(t-\tau+1)}$ in (3), knowing that:

1) if a sensor is attack free (i.e., $b_i = 0$), then (3) reduces to $Y_i^{(t)} - \mathcal{O}_i x^{(t-\tau+1)} = \Psi_i^{(t)}$;
2) $\overline{\Psi}_i$ is the upper bound on the norm of the noise at the ith sensor; and
3) the maximum number of attacked sensors is $\overline{s}$.

Therefore, using the binary variables b_i, we can pose the problem of secure state estimation as follows.

Problem II.1: (Secure State Estimation) For the linear control system under attack defined by (1) and (2), construct an estimate $\eta = (x, b) \in \mathbb{R}^n \times \mathbb{B}^p$ such that $\eta \models \phi$, i.e., η satisfies the formula ϕ:

$$\phi ::= \bigwedge_{i=1}^{p} \left(\neg b_i \Rightarrow \|Y_i - \mathcal{O}_i x\|_2 \leq \overline{\Psi}_i \right) \wedge \left(\sum_{i=1}^{p} b_i \leq \overline{s} \right).$$

The first conjunction of constraints requires $(Y_i - \mathcal{O}_i x)$ to be bounded only by the noise bound if sensor i is attack free. We resort to the 2-norm of $(Y_i - \mathcal{O}_i x)$ since the only information we have available about the noise is a bound on its 2-norm. The second inequality enforces the cardinality constraint on the number of attacked sensors. We use $\models$ to denote that a solution (x, b) satisfies the logic formula ϕ in the problem statement, meaning that the evaluation of ϕ at (x, b) is the Boolean value $\top$ (true). We drop the time argument t in Problem II.1 since the satisfiability problem is to be solved at every time instance.

Problem II.1 does not ask for the minimal number of attacked sensors for which the estimated state matches the measured output. That is, if b^* is the vector of indicator variables characterizing the actual attack, any assignment $\eta = (x, b) \models \phi$ with $\text{supp}(b^*) \subseteq \text{supp}(b)$ is a valid solution for Problem II.1. Therefore, it is useful to modify Problem II.1 to ask for the minimal number of attacked sensors that explains the collected measurements as follows.

Problem II.2: (Minimal Attack Support) For the linear control system under attack defined by (1) and (2), construct the estimate $\eta = (x, b) \in \mathbb{R}^n \times \mathbb{B}^p$ obtained as the solution of the optimization problem:

$$\min_{(x,b) \in \mathbb{R}^n \times \mathbb{B}^p} \sum_{i=1}^{p} b_i \ \text{ s.t. } \bigwedge_{i=1}^{p} \left(\neg b_i \Rightarrow \|Y_i - \mathcal{O}_i x\|_2 \leq \overline{\Psi}_i \right).$$

We observe that a solution for Problem II.2 will also satisfy ϕ and, therefore, is a solution for Problem II.1. In fact, it is straightforward to show that the solution to Problem II.2 can be obtained by performing a binary search over $\overline{s}$ and invoking a solver for Problem II.1 at each step, starting with the maximum value for $\overline{s}$ and then decreasing it until Problem II.1 becomes infeasible or $\overline{s} = 0$. Since any solution of (3) must necessarily satisfy the constraints of Problem II.1, such a procedure will terminate by returning the solution with the minimal attack support. We denote this solution as *minimal support solution*. In the reminder of the paper, we will focus on the analysis of the feasibility Problem II.1, since a solution to the optimization Problem II.2 can be obtained by solving a sequence of instances of Problem II.1.

In Section II-E, we discuss the conditions for the uniqueness of the minimal support solution of Problem II.2. However, we first recall that the satisfiability problem over real numbers, and specifically over $\mathbb{R}^n$, is inherently intractable, i.e., decision algorithms for formulas with nonlinear polynomials already suffer from high complexity [25], [26]. Moreover, linear programming and convex programming solvers usually perform floating point (hence inexact) calculations, which may be inadequate for some applications. Therefore, to provide formal guarantees about the correctness of Problem II.1, we resort to a notion of δ-satisfaction and δ-completeness, inspired by the ones previously proposed by Gao *et al.* [27].

Definition II.3 (Soundness and Completeness of Decision Algorithms for Problem II.1): Let a minimal solution $\eta^* = (x^*, b^*)$ exist for Problem II.2, and hence for Problem II.1 (i.e., $\eta^* \models \phi$), providing the true state and a minimum number of nonzero indicator variables. Then, a solution $\eta = (x, b)$ is said to δ-satisfy ϕ (or δ-SAT for short), denoted by $\eta \models_\delta \phi$, for some $\delta \in \mathbb{R}$, $\delta \geq 0$, if $\text{supp}(b^*) \subseteq \text{supp}(b)$ and $\|x^* - x\|_2^2 \leq \delta$. Moreover, an algorithm that solves Problem II.1 is said to be *δ-complete* if it returns a δ-SAT solution.

Definition II.3 asks for an algorithm that terminates and returns a solution $\eta = (x, b)$ that is correct (up to the tolerance δ). Hence, a *δ-complete* decision algorithm in the sense of Definition II.3 is also (δ-)sound since, if it returns a solution η, η is actually a δ-SAT solution.

E. Uniqueness of Minimal Support Solutions

To characterize the existence and uniqueness of solutions to Problem II.2, we recall the notion of s-sparse observability [16].

Definition II.4: (s-Sparse Observable System) The linear control system under attack defined by (1) and (2) is said to be *s-sparse observable* if for every set $\Gamma \subseteq \{1, \ldots, p\}$ with $|\Gamma| = p - s$, the pair (A, C_Γ) is observable.

In other words, a system is s-sparse observable if it is observable from any choice of $p - s$ sensors. For $2\overline{s}$-sparse observable systems, the following result holds.

Theorem II.5: (Existence and Uniqueness of the Solution)[16, Th. III.2] In the noiseless case ($\Psi_i = 0$ for all $i \in \{1, \dots, p\}$), Problem II.2 admits a unique solution $\eta^* = (x^*, b^*)$ if and only if the dynamical system under attack, defined by (1) and (2), is $2\overline{s}$-sparse observable.

The following result was established as part of the proof of [16, Th. II.5] and will be used in Section III.

Proposition II.6: Let the dynamical system under attack, defined by (1) and (2), be $2\overline{s}$-sparse observable. The observability matrix $\mathcal{O}_{\mathcal{I}}$ has a trivial kernel for any set $\mathcal{I} \subseteq \{1, \dots, p\}$ with $|\mathcal{I}| \geq p - 2\overline{s}$.

Remark II.7: As stated in Theorem II.5, the state of the dynamical system under attack, defined by (1) and (2), can be uniquely determined when the system is $2\overline{s}$-sparse observable. This condition seems expensive to check because of its combinatorial nature: we have to check observability of all possible pairs (A, C_{Γ}). Yet, the $2\overline{s}$-sparse observability condition clearly illustrates a fundamental limitation for secure state estimation: *it is impossible to correctly reconstruct the state whenever a number of sensors larger than or equal to* $\lceil p/2 \rceil$ *is attacked, since there exist different states producing the same observations under the effect of attacks.*

Indeed, suppose that we have an even number of sensors p and $\overline{s} = p/2$ sensors are attacked. Then, Theorem II.5 requires the system to still be observable after removing $2\overline{s} = p$ rows from the map C. However, this is impossible since $C_{\overline{\Gamma}}$ becomes the null matrix. This fundamental limitation is consistent with previous results reported in the literature [7], [28].

Remark II.8: Based on Theorem II.5, the state of the system can be uniquely identified despite the existence of attacks *if and only if* the system is $2\overline{s}$-sparse observable, $\overline{s}$ being an upper bound on the number of attacked sensors. If such a bound $\overline{s}$ is not known *a priori*, we can use Theorem II.5 to determine $\overline{s}$ by removing all the combinations of $2s$ sensors for $s = 1, \dots, \lfloor p/2 \rfloor$ and checking the observability of the resulting system, until we find the maximum possible number of sensors that can be removed while still being able to reconstruct the system state. We observe that such a bound is an intrinsic characteristic of the system, since it only depends on the pair (A, C).

Problem II.2 can be solved by transforming it into a mixed integer-quadratic program (MIQP) as follows:

$$\min_{(x,b) \in \mathbb{R}^n \times \mathbb{B}^p} \sum_{i=1}^{p} b_i \ \text{ s.t. } \ \|Y_i - \mathcal{O}_i x\|_2 \leq M b_i + \overline{\Psi}_i$$
$$1 \leq i \leq p \qquad (5)$$

where $M \in \mathbb{R}$ is a constant that should be "big" enough to make each constraint not active when $b_i = 1$. The relaxation in (5) is typically used to express constraints including logical implications [29]; however, in this case, the choice of M affects the completeness of the approach, which will depend on M. For example, in the absence of noise, since $\|Y_i - \mathcal{O}_i x\|_2$ is ultimately bounded by the power of the attack $\|E_i\|_2$, a value of $M < \|E_i\|_2 = \|Y_i - \mathcal{O}_i x\|_2$, can produce an incorrect result. While a physical sensor has a bounded dynamic range in practice, such a bound is not known *a priori* in our formulation, which makes no assumptions on $\|E_i\|_2$. Therefore, completeness of the MIQP formulation (5) cannot be guaranteed in general.

In the sequel, we detail an algorithm that exploits the geometry of the state estimation problem and the convexity of the quadratic constraints to generate a provably correct solution using the SMT paradigm. We compare the SMT-based solution with the MIQP formulation in (5) using a commercial MIQP solver.

III. SMT-Based Detector

To decide whether a combination of Boolean and convex constraints is satisfiable, we construct the detection algorithm IMHOTEP-SMT using the *lazy* SMT paradigm [21]. As in the CalCS solver [22], our decision procedure combines a SAT solver (SAT-SOLVE) and a theory solver ($\mathcal{T}$-SOLVE) for convex constraints on real numbers. The SAT solver efficiently reasons about combinations of Boolean and pseudo-Boolean constraints, using the David–Putnam–Logemann–Loveland algorithm [30], to suggest possible assignments for the convex constraints. The theory solver checks the consistency of the given assignments, and provides the reason for the conflict, a *certificate*, or a counterexample, whenever inconsistencies are found. Each certificate results in learning new constraints that will be used by the SAT solver to prune the search space. The complex detection and mitigation decision task is thus broken into two simpler tasks, respectively, over the Boolean and convex domains. We denote the approach as lazy, because it checks and learns about consistency of convex constraints only when necessary, as detailed below.

A. Overall Architecture

As illustrated in Algorithm 1, we start by mapping each convex constraint to an auxiliary Boolean variable c_i to obtain the following (pseudo-)Boolean satisfiability problem:

$$\phi_B := \left(\bigwedge_{i \in \{1,\dots,p\}} \neg b_i \Rightarrow c_i \right) \wedge \left(\sum_{i \in \{1,\dots,p\}} b_i \leq \overline{s} \right)$$

where $c_i = 1$ if $\|Y_i - \mathcal{O}_i x\|_2 \leq \overline{\Psi}_i$ is satisfied, and zero otherwise. By only relying on the Boolean structure of the problem, SAT-SOLVE returns an assignment for the variables b_i and c_i (for $i = 1, \dots, p$), thus hypothesizing which sensors are attack free, hence which convex constraints should be jointly satisfied.

This Boolean assignment is then used by $\mathcal{T}$-SOLVE to determine whether there exists a state $x \in \mathbb{R}^n$ that satisfies all the convex constraints related to the unattacked sensors, i.e., $\|Y_i - \mathcal{O}_i x\|_2 \leq \overline{\Psi}_i$ for $i \in \overline{\text{supp}}(b)$. If x is found, IMHOTEP-SMT terminates with `SAT` and provides the solution (x, b). Otherwise, the `UNSAT` certificate ϕ_{cert} is generated in terms of new Boolean constraints, explaining which sensor measurements are conflicting and may be under attack. A very naïve certificate can always be provided in the form of:

$$\phi_{\text{UNSAT-cert}} = \sum_{i \in \overline{\text{supp}}(b)} b_i \geq 1$$

which encodes the fact that at least one of the sensors in the set $\overline{\text{supp}}(b)$ (i.e., for which $b_i = 0$) is actually under attack. The augmented Boolean problem consisting of the original formula

Algorithm 1: IMHOTEP-SMT **Input:** $A, B, C, Y, U, \overline{s}$
Output: $\eta = (x, b)$.

```
status := UNSAT;
$\phi_B := \left(\bigwedge_{i\in\{1,\ldots,p\}} \neg b_i \Rightarrow c_i\right) \wedge \left(\sum_{i\in\{1,\ldots,p\}} b_i \leq \overline{s}\right)$;
while status == UNSAT do
    $(b, c)$ := SAT-SOLVE($\phi_B$);
    (status, $x$) := $\mathcal{T}$-SOLVE.CHECK($\overline{\text{supp}}(b)$);
    if status == UNSAT then
        $\phi_{\text{cert}}$ := $\mathcal{T}$-SOLVE.CERTIFICATE($\overline{\text{supp}}(b), x$);
        $\phi_B := \phi_B \wedge \phi_{\text{cert}}$;
    end if
end while
return $\eta = (x, b)$;
```

ϕ_B and the generated certificate $\phi_{\text{UNSAT-cert}}$ is then fed back to SAT-SOLVE to produce a new assignment. The sequence of new SAT queries is then repeated until $\mathcal{T}$-SOLVE terminates with SAT.

By the $2\overline{s}$-sparse observability condition (see Theorem II.5), there always exists a unique solution to Problem II.2, hence Algorithm 1 will always terminate. While Algorithm 1 is intended to solve Problem II.1, a solution for Problem II.2 can always be obtained, as mentioned earlier, by adding an external loop that conducts a binary search to Algorithm 1, which can increase the overall execution time. However, to help the SAT solver quickly converge toward the correct assignment, a central problem in lazy SMT solving is to generate succinct explanations whenever conjunctions of convex constraints are infeasible, possibly highlighting the minimum set of conflicting assignments. The rest of this section will then focus on the implementation of the two main tasks of $\mathcal{T}$-SOLVE, namely, checking the satisfiability of a given assignment ($\mathcal{T}$-SOLVE.CHECK), and generating succinct UNSAT certificates ($\mathcal{T}$-SOLVE.CERTIFICATE). For clarity's sake, we focus on the noiseless case ($\Psi = 0$) in this section; we will extend our results to the noisy case in Section IV.

B. Satisfiability Checking

Given an assignment of the Boolean variable b, with $|\text{supp}(b)| \leq \overline{s}$, the following condition holds:

$$\min_{x\in\mathbb{R}^n} \left\|Y_{\overline{\text{supp}}(b)} - \mathcal{O}_{\overline{\text{supp}}(b)} x\right\|_2^2 = 0 \tag{6}$$

if and only if $x = x^*$ and $\text{supp}(b) \supseteq \text{supp}(b^*)$, (x^*, b^*) being the solution of Problem II.2. This is a direct consequence of the $2\overline{s}$-sparse observability property discussed in Section II. The preceding *unconstrained least squares optimization* problem can be solved very efficiently, thus leading to Algorithm 2. In practical implementations, (6) should actually be replaced with

$$\min_{x\in\mathbb{R}^n} \left\|Y_{\overline{\text{supp}}(b)} - \mathcal{O}_{\overline{\text{supp}}(b)} x\right\|_2^2 \leq \epsilon$$

where $\epsilon > 0$ is the solver tolerance, accounting for numerical errors. As for the noise, we focus here on the case when ϵ is zero and defer the discussion for nonzero tolerance to the following section.

We characterize the soundness and completeness of Algorithm 2, the basic block of our SMT-based detector, with the following result.

Lemma III.1: Let the linear dynamical system under attack, defined by (1) and (2), be $2\overline{s}$-sparse observable. Let $\overline{\Psi}_i = 0$ for all $i \in \{1, \ldots, p\}$ and let also $\epsilon = 0$ be the numerical solver tolerance for Algorithm 2. Then for any index set $\mathcal{I}$ with cardinality $|\mathcal{I}| \geq p - \overline{s}$, Algorithm 2 returns SAT if and only if the following holds:

1) $\mathcal{I} \subseteq \overline{\text{supp}}(b^*)$,
2) $\|x^* - x\|_2^2 = 0$,

where (x^*, b^*) is the solution to Problem II.2 and x is computed as in line 1 of Algorithm 2.

Proof: Since the "if" condition is trivial to show, we focus on the "only if" condition. Define $\mathcal{I}'$ as the set of indices of the sensors that are attack free. Define also $\mathcal{I}''$ as the set $\mathcal{I}'' = \mathcal{I} \setminus \mathcal{I}'$. We can write the result from lines 1 and 2 of Algorithm 2 as

$$\min_{x\in\mathbb{R}^n} \|Y_{\mathcal{I}} - \mathcal{O}_{\mathcal{I}} x\|_2^2 = 0$$

$$\Rightarrow \min_{x\in\mathbb{R}^n} \sum_{i\in\mathcal{I}} \|Y_i - \mathcal{O}_i x\|_2^2 = 0$$

$$\Rightarrow \min_{x\in\mathbb{R}^n} \sum_{i\in\mathcal{I}'} \|Y_i - \mathcal{O}_i x\|_2^2 + \sum_{i\in\mathcal{I}''} \|Y_i - \mathcal{O}_i x\|_2^2 = 0$$

$$\Rightarrow \min_{x\in\mathbb{R}^n} \|\mathcal{O}_{\mathcal{I}'}(x^* - x)\|_2^2 + \sum_{i\in\mathcal{I}''} \|\mathcal{O}_i(x^* - x) + E_i^*\|_2^2 = 0.$$

Hence, in order for Algorithm 2 to return SAT, both terms $\|\mathcal{O}_{\mathcal{I}'}(x^* - x)\|_2^2$ and $\sum_{i\in\mathcal{I}''} \|\mathcal{O}_i(x^* - x) + E_i^*\|_2^2$ must vanish at the optimal point.

Since at most $\overline{s}$ sensors are under attack, we conclude that $|\mathcal{I}''|$ is at most $\overline{s}$ and $|\mathcal{I}'| \geq p - 2\overline{s}$. Hence, it follows from Proposition II.6 that the observability matrix $\mathcal{O}_{\mathcal{I}'}$ has a trivial kernel. Therefore, we conclude that $\|\mathcal{O}_{\mathcal{I}'}(x^* - x)\|_2^2$ evaluates to zero if and only if $x = x^*$. This, in turn, implies that the solution of the optimization problem in line 1 of Algorithm 2 is x^* and hence $\|x^* - x\|_2^2 = 0$.

To conclude, we need to show that $\mathcal{I} \subseteq \overline{\text{supp}}(b^*)$. However, this follows from the requirement that $\sum_{i\in\mathcal{I}''} \|\mathcal{O}_i(x^* - x) + E_i^*\|_2^2$ vanishes at the optimal point, i.e., for $x = x^*$. Hence

$$\sum_{i\in\mathcal{I}''} \|\mathcal{O}_i(x^* - x) + E_i^*\|_2^2 = 0 \Rightarrow \sum_{i\in\mathcal{I}''} \|E_i^*\|_2^2 = 0$$

which, in turn, implies that all the sensors indexed by $\mathcal{I}''$ are attack free. Combining this result with the definition of the set $\mathcal{I}'$, we conclude that all the sensors indexed by $\mathcal{I}$ are actually attack free, and the inclusion $\mathcal{I} \subseteq \overline{\text{supp}}(b^*)$ holds. ■

When noise or nonzero numerical tolerance is present, we modify Algorithm 2 by checking instead whether the optimal x drives the objective function below the noise level and the numerical tolerance. Clearly, satisfying such a constraint on the 2-norms is not sufficient, in general, to retrieve the actual state in the sense of Definition II.3: attacks having a relatively small power may not be detected. Therefore, in Section IV, we will determine under which conditions on the noise level and the numerical tolerance it is possible to achieve δ-completeness as in Definition II.3.

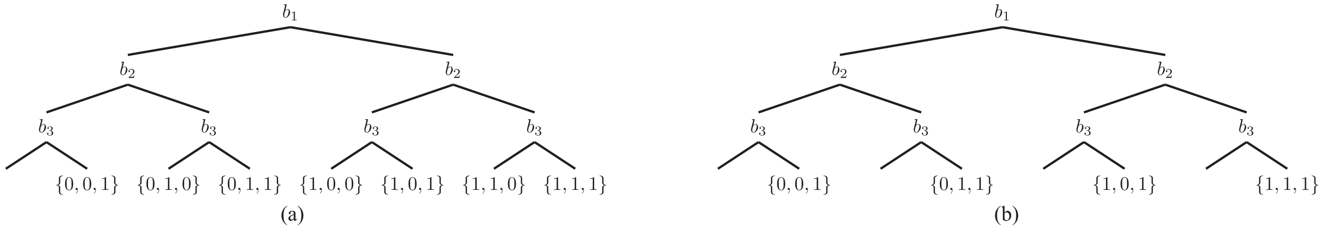


Fig. 1. Pictorial example illustrating the effect of generating smaller conflicting certificates. (a) A tree showing all the combinations of three Boolean indicator variables b_1, b_2, b_3 when a conflicting certificate of the form $\phi_{\text{cert}} := b_1 + b_2 + b_3 \geq 1$ is generated. The missing 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 2: $\mathcal{T}$ -SOLVE.CHECK $(\mathcal{I})$:

```
Solve: $x := \arg\min_{x \in \mathbb{R}^n} \|Y_{\mathcal{I}} - \mathcal{O}_{\mathcal{I}} x\|_2^2$
if $\|Y_{\mathcal{I}} - \mathcal{O}_{\mathcal{I}} x\|_2^2 = 0$ then
    status = SAT;
else
    status = UNSAT;
end if
return (status, $x$);
```

C. Generating Compact UNSAT Certificates

Whenever $\mathcal{T}$-SOLVE.CHECK provides UNSAT, a naïve certificate could be easily generated as mentioned above:

$$\phi_{\text{triv-cert}} = \sum_{i \in \overline{\text{supp}}(b)} b_i \geq 1 \tag{7}$$

indicating that at least one of the sensors, which was initially assumed as attack free (i.e., for which $b_i = 0$), is actually under attack; one of the b_i variables should then be set to one in the next assignment of the SAT solver. However, such *trivial certificate* $\phi_{\text{triv-cert}}$ does not provide much information, since it only excludes the current assignment from the search space, and can lead to exponential execution time, as reflected by the following proposition.

Proposition III.2: Let the linear dynamical system under attack, defined by (1) and (2), be $2\overline{s}$-sparse observable. Let $\overline{\Psi}_i = 0$ for all $i \in \{1, \ldots, p\}$ and let also $\epsilon = 0$ be the numerical solver tolerance for Algorithm 2. Then, Algorithm 1 using the trivial UNSAT certificate $\phi_{\text{triv-cert}}$ in (7) is δ-complete (in the sense of Definition II.3) with $\delta = 0$. Moreover, the upper bound on the number of iterations of Algorithm 1 is $\sum_{s=0}^{\overline{s}} \binom{p}{s}$. ■

Proof: δ-Completeness of Algorithm 1 follows directly from Lemma III.1. To derive the bound on the number of iterations, we first recall that the $2\overline{s}$-sparse observability condition ensures uniqueness of a minimal solution (see Theorem II.5). The worst case scenario would happen when the solver exhaustively explores all possible combinations of attacked sensors with cardinality less than or equal to $\overline{s}$ in order to find the correct assignment. This amounts to $\sum_{s=0}^{\overline{s}} \binom{p}{s}$ iterations.

The generated UNSAT certificates heavily affect the overall execution time of Algorithm 1: the smaller the certificate, the more information is learnt and the faster is the convergence of the SAT solver to the correct assignment. For example, a certificate with $b_i = 1$ would identify exactly one attacked sensor at each step, a substantial improvement with respect to the exponential worst case complexity of the plain SAT problem, which is NP-complete. This intuition is described in Fig. 1, where the effect of generating two certificates with different sizes is shown. Hence, following the approach of CALCS [22], we focus on designing algorithms that can lead to more *compact certificates* to enhance the execution time of IMHOTEP-SMT, by exploiting the specific structure of the secure state estimation problem.

To do so, we first observe that the measurements of each sensor $Y_i = \mathcal{O}_i x$ define an affine subspace $\mathbb{H}_i \subseteq \mathbb{R}^n$ as

$$\mathbb{H}_i = \{x \in \mathbb{R}^n \mid Y_i - \mathcal{O}_i x = 0\}.$$

The dimension of $\mathbb{H}_i$ is given by the dimension of the null space of the matrix $\mathcal{O}_i$, i.e., $\dim(\mathbb{H}_i) = \dim(\ker \mathcal{O}_i)$. Then, satisfiability checking in Algorithm 2 can be reformulated as follows. Let r_i be the *residual* of the state x with respect to the affine subspace $\mathbb{H}_i$, defined as $r_i(x) = \|Y_i - \mathcal{O}_i x\|_2^2$. The optimization problem in Algorithm 2 is equivalent to searching for a point x that minimizes the sum of the individual residuals with respect to all the affine subspaces $\mathbb{H}_i$ for $i \in \mathcal{I}$, i.e.,

$$\min_{x \in \mathbb{R}^n} \|Y_{\mathcal{I}} - \mathcal{O}_{\mathcal{I}} x\|_2^2 = \min_{x \in \mathbb{R}^n} \sum_{i \in \mathcal{I}} \|Y_i - \mathcal{O}_i x\|_2^2 = \min_{x \in \mathbb{R}^n} \sum_{i \in \mathcal{I}} r_i(x).$$

Based on the formulation above, it is straightforward to show the following result.

Proposition III.3: Let the linear dynamical system under attack, defined by (1) and (2), be $2\overline{s}$-sparse observable. Let $\overline{\Psi}_i = 0$ for all $i \in \{1, \ldots, p\}$ and let also $\epsilon = 0$ be the numerical solver tolerance for Algorithm 2. Then, for any set of indices $\mathcal{I} \subseteq \{1, \ldots, p\}$, the following statements are equivalent:

1) $\mathcal{T}$ -SOLVE.CHECK$(\mathcal{I})$ returns UNSAT,
2) $\min_{x \in \mathbb{R}^n} \sum_{i \in \mathcal{I}} r_i(x) > 0$,
3) $\bigcap_{i \in \mathcal{I}} \mathbb{H}_i = \emptyset$.

In the following, we describe two algorithms that can generate two types of compact certificates, namely conflicting certificates and agreeable certificates.

D. Certificate Based on Smaller Conflicting Sensor Sets

To generate a compact Boolean constraint that explains a conflict, we aim to find a small set of sensors that cannot all be attack free. A key result of this paper is to show that such set exists and can be computed in time that is linear in the size of the problem. This is captured by the following proposition whose proof exploits the geometric interpretation provided by the affine subspaces $\mathbb{H}_i$.

Lemma III.4: Let the linear dynamical system under attack, defined by (1) and (2), be $2\overline{s}$-sparse observable. Let $\overline{\Psi}_i = 0$ for all $i \in \{1, \ldots, p\}$ and let also $\epsilon = 0$ be the numerical solver tolerance for Algorithm 2. If $\mathcal{T}$-SOLVE.CHECK$(\mathcal{I})$ is UNSAT for a set $\mathcal{I}$, with $|\mathcal{I}| > p - 2\overline{s}$, then there exists a subset $\mathcal{I}_{\text{temp}} \subset \mathcal{I}$ with $|\mathcal{I}_{\text{temp}}| \leq p - 2\overline{s} + 1$ such that $\mathcal{T}$-SOLVE.CHECK$(\mathcal{I}_{\text{temp}})$ is also UNSAT. Moreover, the time complexity of finding $\mathcal{I}_{\text{temp}}$ is linear in both p and $\overline{s}$.

Proof: Consider any set of sensors $\mathcal{I}' \subset \mathcal{I}$ such that $|\mathcal{I}'| = p - 2\overline{s}$ and $\bigcap_{i\in\mathcal{I}'} \mathbb{H}_i$ is not empty. If such set $\mathcal{I}'$ does not exist, then the result follows trivially. If the set $\mathcal{I}'$ exists, then it follows from Proposition II.6 that $\mathcal{O}_{\mathcal{I}'}$ has a trivial kernel and hence the intersection $\bigcap_{i\in\mathcal{I}'} \mathbb{H}_i$ is a single point, named x'. Now, since $\mathcal{T}$-SOLVE.CHECK$(\mathcal{I})$ is UNSAT, it follows from Proposition III.3 that:

$$\bigcap_{i\in\mathcal{I}} \mathbb{H}_i = \emptyset \Rightarrow \bigcap_{i\in\mathcal{I}'} \mathbb{H}_i \cap \bigcap_{i\in\mathcal{I}\setminus\mathcal{I}'} \mathbb{H}_i = \emptyset \Rightarrow \{x'\} \cap \bigcap_{i\in\mathcal{I}\setminus\mathcal{I}'} \mathbb{H}_i = \emptyset$$

which in turn implies that there exists at least one sensor $i \in \mathcal{I} \setminus \mathcal{I}'$ such that its affine subspace $\mathbb{H}_i$ does not pass through the point x'. Now, we define $\mathcal{I}_{\text{temp}}$ as $\mathcal{I}_{\text{temp}} = \mathcal{I}' \cup i$ and we note that $|\mathcal{I}_{\text{temp}}| = p - 2\overline{s} + 1$, which is what we wanted to show. To conclude the proof, the linear complexity follows from the construction in Algorithm 3 detailed below. ■

Using Lemma III.4, our objective is to find a small set of affine subspaces that fail to intersect. Based on the intuition in the proof of Lemma III.4, our algorithm works as follows. First, we construct the set of indices $\mathcal{I}'$ by picking any random set of $p - 2\overline{s}$ sensors. We then search for one additional sensor i, which can lead to a conflict with the sensors indexed by $\mathcal{I}'$. To do this, we call $\mathcal{T}$-SOLVE.CHECK by passing the set $\mathcal{I}_{\text{temp}} := \mathcal{I}' \cup i$ as an argument. If the check returns SAT, then we label these sensors as "nonconflicting" and we repeat the same process by replacing the sensor indexed by i with another sensor until we reach a conflicting set of affine subspaces. Termination of this process is guaranteed by Lemma III.4, thus revealing a set of $p - 2\overline{s} + 1$ conflicting affine subspaces. Once the set is discovered, we stop by generating the following, more compact, certificate:

$$\phi_{\text{conf-cert}} := \sum_{i\in\mathcal{I}_{\text{temp}}} b_i \geq 1.$$

These steps are summarized in Algorithm 3. While Algorithm 3 is guaranteed to terminate regardless of the initial random set $\mathcal{I}'$ or the order in which the sensor i is selected, the execution time may change. In Algorithm 4, we show the heuristics used to implement the two steps of Algorithm 3, namely, the selection of the initial set $\mathcal{I}'$ and the further addition of sensor indexes, which further exploit the geometry of our problem.

Our conjecture is that the $p - 2\overline{s}$ affine subspaces with the lowest (normalized) residuals are most likely to have a common intersection point, which can then be used as a candidate intersection point for the affine subspaces against the higher (normalized) residuals, one-by-one, until a conflict is detected. A pictorial illustration of this intuition is given in Fig. 2(a). Based on this intuition, we first compute the (normalized) residuals r_i for all $i \in \mathcal{I}$, and sort them in ascending order. We then

Algorithm 3: $\mathcal{T}$-SOLVE.CERTIFICATE-CONFLICT-ORIG$(\mathcal{I}, x)$.

```
Step 1: Pick a set $\mathcal{I}' \subset \mathcal{I}$ of $p - 2\overline{s}$ sensors;
Step 2: Conduct a linear search for the UNSAT certificate:
status = SAT;
Pick a sensor index $i \in \mathcal{I} \setminus \mathcal{I}'$;
    $\mathcal{I}\_temp := \mathcal{I}' \cup i$;
while status == SAT do
  (status, $x$) := $\mathcal{T}$-SOLVE.CHECK($\mathcal{I}\_temp$);
  if status == UNSAT then
    $\phi_{\text{conf-cert}} := \sum_{i\in\mathcal{I}\_temp} b_i \geq 1$;
  else
    Pick another sensor index $i \in \mathcal{I} \setminus \mathcal{I}'$;
    $\mathcal{I}\_temp := \mathcal{I}' \cup i$;
  end if
end while
return $\phi_{\text{conf-cert}}$;
```

Algorithm 4: $\mathcal{T}$-SOLVE.CERTIFICATE-CONFLICT$(\mathcal{I}, x)$.

```
Compute normalized residuals
  $r := \bigcup_{i\in\mathcal{I}} \{r_i\},\ r_i := \|Y_i - \mathcal{O}_i x\|_2^2 \,/\, \|\mathcal{O}_i\|_2^2,\ i \in \mathcal{I}$;
Sort the residual variables
  $r\_sorted$ := sortAscendingly($r$);
Pick the index corresponding to the maximum residual
  $\mathcal{I}\_max\_r$ := Index($r\_sorted_{\{|\mathcal{I}|,|\mathcal{I}|-1,\ldots,p-2\overline{s}+1\}}$);
  $\mathcal{I}\_min\_r$ := Index($r\_sorted_{\{1,\ldots,p-2\overline{s}\}}$);
Search linearly for the UNSAT certificate
  status = SAT;   counter = 1;
  $\mathcal{I}\_temp := \mathcal{I}\_min\_r \cup \mathcal{I}\_max\_r_{counter}$;
while status == SAT do
  (status, $x$) := $\mathcal{T}$-SOLVE.CHECK($\mathcal{I}\_temp$);
  if status == UNSAT then
    $\phi_{\text{conf-cert}} := \sum_{i\in\mathcal{I}\_temp} b_i \geq 1$;
  else
    counter := counter + 1;
    $\mathcal{I}\_temp := \mathcal{I}\_min\_r \cup \mathcal{I}\_max\_r_{counter}$;
  end if
end while
[Optional] Sort the rest according to dim($ker\{\mathcal{O}\}$)
  $\mathcal{I}\_temp2$ = sortAscendingly($dim(ker\{\mathcal{O}_{\mathcal{I}\_temp}\})$);
  status = UNSAT;   counter2 = $|\mathcal{I}\_temp2| - 1$;
  $\mathcal{I}\_temp2 := \mathcal{I}\_temp2_{\{1,\ldots,counter2\}}$;
while status == UNSAT do
  (status, $x$) := $\mathcal{T}$-SOLVE.CHECK($\mathcal{I}_{\text{temp}}$);
  if status == SAT then
    $\phi_{\text{conf-cert}} := \sum_{i\in\mathcal{I}\_temp2_{\{1,\ldots,counter2+1\}}} b_i \geq 1$;
  else
    counter2 := counter2 - 1;
    $\mathcal{I}\_temp2 := \mathcal{I}\_temp2_{\{1,\ldots,counter2\}}$;
  end if
end while
return $\phi_{\text{conf-cert}}$
```

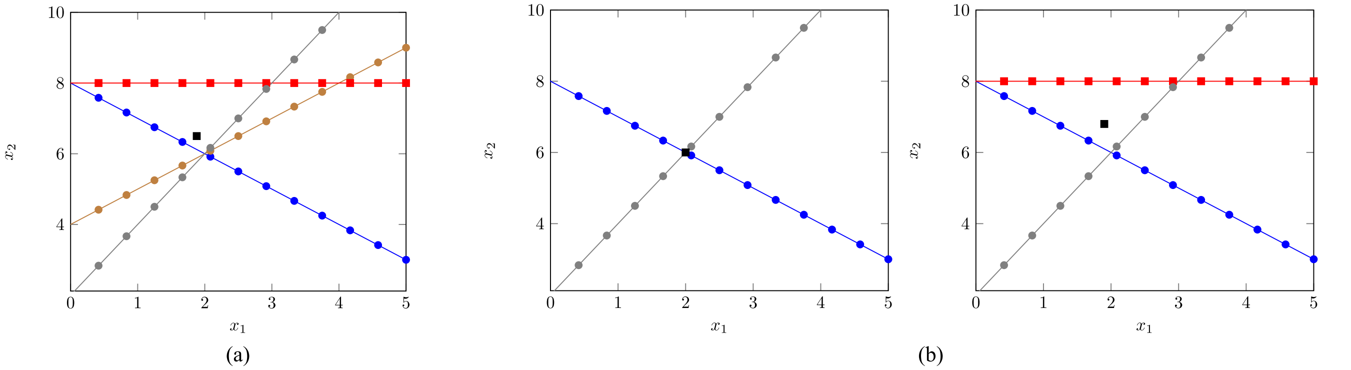


Fig. 2. Pictorial examples illustrating the geometrical intuitions behind Algorithm 4. (a) Four affine subspaces corresponding to measurements from four different sensors. The red affine subspace corresponds to the sensor under attack. All other affine subspaces intersect at the unique solution. The optimal point is marked as a black box. (b) An example of a run of Algorithm 4. In the first iteration (left), the set $\mathcal{I}\_min\_r$ contains the $p - 2\overline{s} = 4 - 2 \times 1 = 2$ indexes of the sensors that correspond to the minimal residuals. This set is a nonconflicting set and hence the corresponding affine subspaces have a unique intersection point. In the second iteration (right), the index of the sensor corresponding to the maximum residual is added to the set $\mathcal{I}_{\text{temp}}$ resulting into a conflict. Algorithm 4 terminates and returns the conflicting set $\mathcal{I}_{\text{temp}}$. In both cases, the optimal point is marked as a black box.

pick the $p - 2\overline{s}$ minimum (normalized) residuals indexed by $\mathcal{I}\_min\_r$, and search for one more affine subspace that leads to a conflict with the affine subspaces indexed by $\mathcal{I}\_min\_r$. To do this, we start by solving the same optimization problem as in Algorithm 2, but on the reduced set of affine subspaces indexed by $\mathcal{I}_{\text{temp}} = \mathcal{I}\_min\_r \cup \mathcal{I}\_max\_r$, where $\mathcal{I}\_max\_r$ is the index associated with the affine subspace having the maximal (normalized) residual. If this set of affine subspaces intersect in one point, they are labeled as "nonconflicting," and we repeat the same process by replacing the affine subspace indexed by $\mathcal{I}\_max\_r$ with the affine subspace associated with the second maximal (normalized) residual from the sorted list, till we reach a conflicting set of affine subspaces. Once the set is discovered, we stop and generate the compact certificate using the sensors indexed in $\mathcal{I}_{\text{temp}}$. A sample execution of Algorithm 4 is illustrated in Fig. 2(b).

Finally, as a postprocessing step, we can further reduce the cardinality of $\mathcal{I}_{\text{temp}}$ by exploiting the dimension of the affine subspaces corresponding to the index list. Intuitively, the lower the dimension, the more information is provided by the corresponding sensor. For example, a sensor i with $\dim(\mathbb{H}_i) = \dim(\ker \mathcal{O}_i) = 0$ can be used to uniquely reconstruct the state. This restricts the search space to the unique point and makes it easier to generate a conflict formula. Therefore, to converge faster toward a conflict, we iterate through the indexes in $\mathcal{I}_{\text{temp}}$ and remove at each step the one which corresponds to the affine subspace with the highest dimension until we are left with a reduced index set that is still conflicting. The following result provides an upper bound for the performance of the proposed heuristics.

Proposition III.5: Let the linear dynamical system under attack, defined by (1) and (2), be $2\overline{s}$-sparse observable. Let $\overline{\Psi}_i = 0$ for all $i \in \{1, \ldots, p\}$ and let also $\epsilon = 0$ be the numerical solver tolerance for Algorithm 2. Then, Algorithm 1 using the conflicting UNSAT certificate $\phi_{\text{conf-cert}}$ in Algorithm 4 is δ-complete (in the sense of Definition II.3) with $\delta = 0$. Moreover, the upper bound on the number of iterations of Algorithm 1 is $\binom{p}{p-2\overline{s}+1}$.

Proof: δ-Completeness follows from Lemma III.1 along with the $2\overline{s}$ observability condition. The upper bound on the number of iterations of Algorithm 1 can be derived as follows. First, it follows from Lemma III.4 that each certificate $\phi_{\text{conf-cert}}$ has at most $p - 2\overline{s} + 1$ sensors. Since we know that the algorithm always terminates, the worst case would then happen when the solver exhaustively generates all conflicting sets of cardinality $p - 2\overline{s} + 1$. This leads to a number of iterations equal to $\binom{p}{p-2\overline{s}+1}$. ■

E. Certificate Based on Agreeable Sensor Sets

To further enhance the solver runtime, we design an algorithm that aims to find a set of $p - 2\overline{s}$ sensors that all agree on the same x. We recall that the $2\overline{s}$-sparse observability condition ensures that the state is fully observable from any set of $p - 2\overline{s}$ sensors. Accordingly, for a given set of sensors, we select the $p - 2\overline{s}$ sensors, hence affine subspaces, that correspond to minimal residuals. We then check whether they all intersect in one point x. In such case, we inform the SAT solver that all of these sensors are unattacked, by generating the following certificate:

$$\phi_{\text{agree-cert}} := \sum_{i \in \mathcal{I}\_min\_r} b_i = 0$$

where $\mathcal{I}\_min\_r$ is the set of indexes of the $p - 2\overline{s}$ affine subspaces with the lowest residuals.

The procedure described above is summarized in Algorithm 5. As evident from line 9 of Algorithm 5, $\phi_{\text{agree-cert}}$ is not always generated; therefore, we use this heuristic, when it is successful, only as a complement of the previously discussed UNSAT certificate. Moreover, the heuristic itself is not always applicable. In fact, it is still possible to design an attack such that up to $\overline{s}$ attacked sensors agree on a single value of x. Hence, unlike our previous results, a stricter assumption of $3\overline{s}$-sparse observability is required, as detailed in the following proposition.

Proposition III.6: Let the linear dynamical system under attack, defined by (1) and (2), be $3\overline{s}$-sparse observable. Let $\overline{\Psi}_i = 0$ for all $i \in \{1, \ldots, p\}$ and $\epsilon = 0$ be the numerical solver tolerance for Algorithm 2. Then, Algorithm 1 using the agreeable UNSAT certificate $\phi_{\text{agree-cert}}$ in Algorithm 5 is δ-

Algorithm 5: $\mathcal{T}$-SOLVE.CERTIFICATE-AGREE$(\mathcal{I}, x)$.

```
Compute normalized residuals
  $r := \bigcup_{i\in\mathcal{I}} \{r_i\}\,,\ r_i := \|Y_i - \mathcal{O}_i x\|_2^2 \,/\, \|\mathcal{O}_i\|_2^2\,,\ i \in \mathcal{I};$
Sort the residual variables
  $r\_sorted :=$ sort Ascendingly$(r)$;
Pick the $p - 2\overline{s}$ indexes corresponding to the minimum residuals
  $\mathcal{I}\_min\_r :=$ Index$(r\_sorted_{\{1,\ldots,p-2\overline{s}\}})$);
  (status, $x$) := $\mathcal{T}$-SOLVE.CHECK$(\mathcal{I}\_min\_r)$;
$\phi_{\text{agree-cert}} :=$ TRUE;
if status == SAT then
  $\phi_{\text{agree-cert}} := \sum_{i\in\mathcal{I}\_min\_r} b_i = 0$;
end if
return $\phi_{\text{agree-cert}}$
```

Algorithm 6: $\mathcal{T}$-SOLVE.CERTIFICATE$(\mathcal{I}, x)$.

```
$\phi_{\text{cert}} := \mathcal{T}$-SOLVE.CERTIFICATE-CONFLICT$(\mathcal{I}, x)$;
if $p > 3\overline{s}$ then
  $\phi_{\text{agree-cert}} := \mathcal{T}$-SOLVE.CERTIFICATE-AGREE$(\mathcal{I}, x)$;
  $\phi_{\text{cert}} := \phi_{\text{cert}} \wedge \phi_{\text{agree-cert}}$;
end if
return $\phi_{\text{cert}}$
```

complete (in the sense of Definition II.3) with $\delta = 0$. Moreover, whenever $\phi_{\text{agree-cert}}$ is generated, Algorithm 1 terminates within $\sum_{s\,=\,0}^{\overline{s}} \binom{2\overline{s}}{s}$ iterations.

Proof: δ-Completeness of Algorithm 1 is equivalent to showing the soundness and completeness of Algorithm 2. It follows from Proposition III.1 that Algorithm 2 is sound and complete whenever the system is $2\overline{s}$-sparse observable and when the cardinality of $\mathcal{I}$ satisfies $|\mathcal{I}| \geq p - \overline{s}$. Hence, to show the result, it is enough to replicate the proof of Proposition III.1 under the assumption that the system is $3\overline{s}$-sparse observable and the cardinality of $\mathcal{I}$ satisfies instead $|\mathcal{I}| \geq p - 2\overline{s}$.

The bound on the number of iterations can be derived as follows. First, we note that $\phi_{\text{agree-cert}}$ assigns $p - 2\overline{s}$ as being unattacked sensors. This in turn forces the solver to search for the attacked sensors in the remaining set of sensors with caridinality $p - (p - 2\overline{s}) = 2\overline{s}$. The bound then follows using the same argument of Proposition III.2. ■

F. Soundness and Completeness of Algorithm 1 in the Noiseless Case

The procedure $\mathcal{T}$-SOLVE.CERTIFICATE$(\mathcal{I}, x)$ in line 7 of Algorithm 1 can be implemented as shown in Algorithm 6. We are now ready to state the main result of this section, which is a direct consequence of our previous results.

Theorem III.7: Let the linear dynamical system under attack, defined by (1) and (2), be $2\overline{s}$-sparse observable, $\overline{\Psi}_i = 0$ for all $i \in \{1, \ldots, p\}$, and $\epsilon = 0$ be the numerical solver tolerance for Algorithm 2. Algorithm 1 is δ-complete (in the sense of Definition II.3) with $\delta = 0$.

IV. COMPLETENESS IN THE PRESENCE OF NOISE

As discussed in the previous section, IMHOTEP-SMT can always detect any compromised sensors in the absence of measurement noise ($\overline{\Psi}_i = 0$ for all $i \in \{1, \ldots, p\}$) and when the numerical tolerance is zero ($\epsilon = 0$). In this section, we characterize completeness in the presence of noise or numerical errors in the solver, by determining to what extent an attack signal can be hidden by noise or the numerical tolerance, thereby making it infeasible to reconstruct the true state. Since Algorithm 1 consists of multiple invocations of the least squares problem, the completeness of the detector entirely depends on the correctness of Algorithm 2 in checking the satisfiability of a Boolean assignment over b.

The completeness of Algorithm 2 will in turn depend on two major components: the tolerance of the numerical solvers, which is typically a small value used as a stopping criterion, and can be controlled by the user; and the noise margin intrinsic to the dynamical system model. To account for these two components, we replace the satisfiability condition in line 2 of Algorithm 2 with the following condition:

$$\|Y_{\mathcal{I}} - \mathcal{O}_{\mathcal{I}} x\|_2 \leq \overline{\Psi}_{\mathcal{I}} + \epsilon \tag{8}$$

where $\epsilon > 0$ is the user-defined tolerance. Then, we recall that the solution of the unconstrained least squares problem in Algorithm 2 is given by

$$x = \left(\mathcal{O}_{\mathcal{I}}^T \mathcal{O}_{\mathcal{I}}\right)^{-1} \mathcal{O}_{\mathcal{I}}^T Y_{\mathcal{I}} = \mathcal{O}_{\mathcal{I}}^{+}\, Y_{\mathcal{I}}$$

where $\mathcal{O}_{\mathcal{I}}^{+} = \left(\mathcal{O}_{\mathcal{I}}^T \mathcal{O}_{\mathcal{I}}\right)^{-1} \mathcal{O}_{\mathcal{I}}^T$ is the Moore–Penrose pseudoinverse of $\mathcal{O}_{\mathcal{I}}$. It is apparent that soundness and completeness of Algorithm 2 depends on the properties of the matrix $\mathcal{O}_{\mathcal{I}}^{+}$. Accordingly, we define the following two quantities.

Definition IV.1: Define $\overline{o} \in \mathbb{R}^+$ as

$$\overline{o} = \max_{\substack{\mathcal{I}\subseteq\{1,\ldots,p\},\\ |\mathcal{I}|\geq p-\overline{s}}} \left\|\mathcal{O}_{\mathcal{I}}^{+}\right\|_2^2$$

where $\mathcal{O}_{\mathcal{I}}^{+}$ is the Moore–Penrose pseudoinverse of $\mathcal{O}_{\mathcal{I}}$.

Definition (Proposition) IV.2: Let the linear system defined in (1) be $2\overline{s}$-sparse observable and define $\Delta_s \in \mathbb{R}^+$ as

$$\Delta_s = \max_{\substack{\Gamma\subset\mathcal{I}\subseteq\{1,\ldots,p\}\\ |\Gamma|\leq\overline{s},|\mathcal{I}|\geq p-\overline{s}}} \lambda_{\max}\left\{\left(\sum_{i\in\Gamma}\mathcal{O}_i^T\mathcal{O}_i\right)\left(\sum_{i\in\mathcal{I}}\mathcal{O}_i^T\mathcal{O}_i\right)^{-1}\right\}.$$

Then, for any $\overline{s}$-sparse attack vector E, and any set $\mathcal{I} \subseteq \{1, \ldots, p\}$, with $|\mathcal{I}| \geq p - \overline{s}$, the following holds:

$$\left\|(I - \mathcal{O}_{\mathcal{I}}\mathcal{O}_{\mathcal{I}}^{+})E_{\mathcal{I}}\right\|_2^2 \geq (1 - \Delta_s)\,\|E_{\mathcal{I}}\|_2^2$$

with Δ_s strictly less than 1.

Proof: We first define the set $\Gamma^* \subset \mathcal{I}$ as the set of indices on which the attack vector E is supported, and note that $E_{\overline{\Gamma^*}} = 0$. Hence:

$$\begin{aligned}\left\|(I - \mathcal{O}_{\mathcal{I}}\mathcal{O}_{\mathcal{I}}^{+})E_{\mathcal{I}}\right\|_2^2 &= E_{\mathcal{I}}^T\left(I - \mathcal{O}_{\mathcal{I}}\mathcal{O}_{\mathcal{I}}^{+}\right)^2 E_{\mathcal{I}}\\ &\overset{(a)}{=} E_{\mathcal{I}}^T\left(I - \mathcal{O}_{\mathcal{I}}\mathcal{O}_{\mathcal{I}}^{+}\right) E_{\mathcal{I}}\\ &= E_{\mathcal{I}}^T E_{\mathcal{I}} - E_{\mathcal{I}}^T \mathcal{O}_{\mathcal{I}}(\mathcal{O}_{\mathcal{I}}^T\mathcal{O}_{\mathcal{I}})^{-1}\mathcal{O}_{\mathcal{I}}^T E_{\mathcal{I}}\\ &\overset{(b)}{=} E_{\Gamma^*}^T E_{\Gamma^*} - E_{\Gamma^*}^T \mathcal{O}_{\Gamma^*}(\mathcal{O}_{\mathcal{I}}^T\mathcal{O}_{\mathcal{I}})^{-1}\mathcal{O}_{\Gamma^*}^T E_{\Gamma^*}\end{aligned} \tag{9}$$

where equality (a) follows from the fact that the matrix $I - \mathcal{O}_{\mathcal{I}}\mathcal{O}_{\mathcal{I}}^{+}$ is idempotent and equality (b) follows from the definition of the set Γ^*. The second term on the right-hand side of the equality (9) can be bounded as

$$E_{\Gamma^*}^T \mathcal{O}_{\Gamma^*} (\mathcal{O}_{\mathcal{I}}^T \mathcal{O}_{\mathcal{I}})^{-1} \mathcal{O}_{\Gamma^*}^T E_{\Gamma^*} \leq \lambda_{\max}\{\mathcal{O}_{\Gamma^*}(\mathcal{O}_{\mathcal{I}}^T \mathcal{O}_{\mathcal{I}})^{-1}\mathcal{O}_{\Gamma^*}^T\} E_{\Gamma^*}^T E_{\Gamma^*}. \tag{10}$$

Moreover, we recall that for any two matrices A and B with appropriate dimensions, $\lambda_{\max}\{AB\} = \lambda_{\max}\{BA\}$. Hence, we can rewrite the right-hand side of (10) as

$$\lambda_{\max}\{\mathcal{O}_{\Gamma^*}^T \mathcal{O}_{\Gamma^*}(\mathcal{O}_{\mathcal{I}}^T \mathcal{O}_{\mathcal{I}})^{-1}\} E_{\Gamma^*}^T E_{\Gamma^*}.$$

Finally, to show that Δ_s is strictly less than one, we recall that $\lambda_{\max}\{\mathcal{O}_{\Gamma^*}^T \mathcal{O}_{\Gamma^*}(\mathcal{O}_{\mathcal{I}}^T \mathcal{O}_{\mathcal{I}})^{-1}\} \leq \Delta_s$ by definition and the equality is achievable. Therefore, it is sufficient to show that the inequality

$$\lambda_{\max}\{\mathcal{O}_{\Gamma}^T \mathcal{O}_{\Gamma}(\mathcal{O}_{\mathcal{I}}^T \mathcal{O}_{\mathcal{I}})^{-1}\} < 1 \tag{11}$$

holds for any set $\mathcal{I}$ and $\Gamma \subset \mathcal{I}$ with $|\Gamma| \leq \overline{s}$ and $|\mathcal{I}| \geq p - \overline{s}$. For this purpose, we notice that

$$\mathcal{O}_{\mathcal{I}}^T \mathcal{O}_{\mathcal{I}} = \sum_{i \in \mathcal{I}} \mathcal{O}_i^T \mathcal{O}_i = \sum_{i \in \Gamma} \mathcal{O}_i^T \mathcal{O}_i + \sum_{i \in \mathcal{I} \setminus \Gamma} \mathcal{O}_i^T \mathcal{O}_i$$
$$= \mathcal{O}_{\Gamma}^T \mathcal{O}_{\Gamma} + \mathcal{O}_{\mathcal{I}\setminus\Gamma}^T \mathcal{O}_{\mathcal{I}\setminus\Gamma}$$

and rewrite (11) as

$$\lambda_{\max}\left\{\mathcal{O}_{\Gamma}^T \mathcal{O}_{\Gamma}\left(\mathcal{O}_{\Gamma}^T \mathcal{O}_{\Gamma} + \mathcal{O}_{\mathcal{I}\setminus\Gamma}^T \mathcal{O}_{\mathcal{I}\setminus\Gamma}\right)^{-1}\right\} < 1$$

where the set $\mathcal{I} \setminus \Gamma$ has a cardinality of at least $p - 2\overline{s}$. Hence, it follows from the $2\overline{s}$-sparse observability condition that the matrix $\mathcal{O}_{\mathcal{I}\setminus\Gamma}^T \mathcal{O}_{\mathcal{I}\setminus\Gamma}$ is positive definite and therefore we can apply Proposition VI.1 in the appendix to show that the statement holds. ■

Using the two quantities defined above, we can state our main result, which is the version of Theorem III.7 in the presence of noise.

Theorem IV.3: Let the linear system defined in (1) be $2\overline{s}$-sparse observable, let $\epsilon > 0$ be the numerical solver tolerance. Then, if each attack signal E_i, for all $i \in S \subset \{1, \ldots, p\}$ with $|S| \leq \overline{s}$, satisfies

$$\|E_i\|_2 > \left(\frac{2}{\sqrt{1-\Delta_s}}\right)\overline{\Psi} + \frac{\sqrt{\epsilon}}{\sqrt{1-\Delta_s}} \tag{12}$$

then Algorithm 1, modified as in (8), is δ-complete with $\delta = \overline{o}\overline{\Psi}^2$.

Proof: To prove the result, we need to show that the condition (8), resulting in δ-satisfiability, is satisfied if and only if no sensor in $\mathcal{I}$ is under attack. If no sensor is under attack, condition (8) is trivially satisfied. Therefore, we focus on proving the reverse implication, showing that if at least one sensor $i_a \in \mathcal{I}$ is under attack, then (8) does not hold as long as the attack E_{i_a} satisfies (12).

To do so, we consider the set $\mathcal{I}$ that contains the attacked sensor i_a, and recall that the solution of the unconstrained least squares problem in Algorithm 2 is given by

$$x = \left(\mathcal{O}_{\mathcal{I}}^T \mathcal{O}_{\mathcal{I}}\right)^{-1} \mathcal{O}_{\mathcal{I}}^T Y_{\mathcal{I}} = \mathcal{O}_{\mathcal{I}}^{+} Y_{\mathcal{I}}$$

where $\mathcal{O}_{\mathcal{I}}^{+} = \left(\mathcal{O}_{\mathcal{I}}^T \mathcal{O}_{\mathcal{I}}\right)^{-1} \mathcal{O}_{\mathcal{I}}^T$ is the Moore–Penrose pseudoinverse of $\mathcal{O}_{\mathcal{I}}$. Hence, the value of the objective function at the optimal point x can be bounded from below as

$$\begin{aligned}
\|Y_{\mathcal{I}} - \mathcal{O}_{\mathcal{I}} x\|_2^2 &\overset{(a)}{=} \left\|Y_{\mathcal{I}} - \mathcal{O}_{\mathcal{I}}\mathcal{O}_{\mathcal{I}}^{+} Y_{\mathcal{I}}\right\|_2^2 \\
&\overset{(b)}{=} \left\|(I - \mathcal{O}_{\mathcal{I}}\mathcal{O}_{\mathcal{I}}^{+})(\mathcal{O}_{\mathcal{I}} x^* + \Psi_{\mathcal{I}} + E_{\mathcal{I}})\right\|_2^2 \\
&\overset{(c)}{=} \left\|(I - \mathcal{O}_{\mathcal{I}}\mathcal{O}_{\mathcal{I}}^{+})(\Psi_{\mathcal{I}} + E_{\mathcal{I}})\right\|_2^2 \\
&\overset{(d)}{\geq} \left(\left\|(I - \mathcal{O}_{\mathcal{I}}\mathcal{O}_{\mathcal{I}}^{+})E_{\mathcal{I}}\right\|_2 - \left\|(I - \mathcal{O}_{\mathcal{I}}\mathcal{O}_{\mathcal{I}}^{+})\Psi_{\mathcal{I}}\right\|_2\right)^2
\end{aligned} \tag{13}$$

where (a) follows from the definition of x, (b) and (c) follow from the definition of $Y_{\mathcal{I}}$ as in (3) and the fact that $\mathcal{O}_{\mathcal{I}}\mathcal{O}_{\mathcal{I}}^{+} \mathcal{O}_{\mathcal{I}} = \mathcal{O}_{\mathcal{I}}$. Finally, the inequality in (d) follows from the inverse triangular inequality.

On the other hand, the condition on the attack signal (12) implies that:

$$\begin{aligned}
\|E_{i_a}\|_2 &> \left(\frac{2}{\sqrt{1-\Delta_s}}\right)\overline{\Psi} + \frac{\sqrt{\epsilon}}{\sqrt{1-\Delta_s}} \\
&\geq \left(\frac{2}{\sqrt{1-\Delta_s}}\right)\overline{\Psi}_{\mathcal{I}} + \frac{\sqrt{\epsilon}}{\sqrt{1-\Delta_s}}.
\end{aligned}$$

Therefore, by noticing that $\|E_{\mathcal{I}}\|_2^2 \geq \|E_{i_a}\|_2^2$ since $i_a \in \mathcal{I}$, we conclude that

$$\begin{aligned}
&\|E_{\mathcal{I}}\|_2 > \left(\frac{2}{\sqrt{1-\Delta_s}}\right)\overline{\Psi}_{\mathcal{I}} + \frac{\sqrt{\epsilon}}{\sqrt{1-\Delta_s}} \\
&\Rightarrow \sqrt{(1-\Delta_s)}\,\|E_{\mathcal{I}}\|_2 > \overline{\Psi}_{\mathcal{I}} + \overline{\Psi}_{\mathcal{I}} + \sqrt{\epsilon} \\
&\overset{(e)}{\Rightarrow} \sqrt{(1-\Delta_s)}\,\|E_{\mathcal{I}}\|_2 > \overline{\Psi}_{\mathcal{I}} + \left\|I - \mathcal{O}_{\mathcal{I}}\mathcal{O}_{\mathcal{I}}^{+}\right\|_2 \|\Psi_{\mathcal{I}}\|_2 + \sqrt{\epsilon} \\
&\overset{(f)}{\Rightarrow} \sqrt{(1-\Delta_s)}\,\|E_{\mathcal{I}}\|_2 > \overline{\Psi}_{\mathcal{I}} + \left\|(I - \mathcal{O}_{\mathcal{I}}\mathcal{O}_{\mathcal{I}}^{+})\Psi_{\mathcal{I}}\right\|_2 + \sqrt{\epsilon} \\
&\overset{(g)}{\Rightarrow} \left\|(I - \mathcal{O}_{\mathcal{I}}\mathcal{O}_{\mathcal{I}}^{+})E_{\mathcal{I}}\right\|_2 > \overline{\Psi}_{\mathcal{I}} + \left\|(I - \mathcal{O}_{\mathcal{I}}\mathcal{O}_{\mathcal{I}}^{+})\Psi_{\mathcal{I}}\right\|_2 + \sqrt{\epsilon} \\
&\Rightarrow \left\|(I - \mathcal{O}_{\mathcal{I}}\mathcal{O}_{\mathcal{I}}^{+})E_{\mathcal{I}}\right\|_2 - \left\|(I - \mathcal{O}_{\mathcal{I}}\mathcal{O}_{\mathcal{I}}^{+})\Psi_{\mathcal{I}}\right\|_2 > \overline{\Psi}_{\mathcal{I}} + \sqrt{\epsilon} \\
&\overset{(h)}{\Rightarrow} \left(\left\|(I - \mathcal{O}_{\mathcal{I}}\mathcal{O}_{\mathcal{I}}^{+})E_{\mathcal{I}}\right\|_2 - \left\|(I - \mathcal{O}_{\mathcal{I}}\mathcal{O}_{\mathcal{I}}^{+})\Psi_{\mathcal{I}}\right\|_2\right)^2 > \overline{\Psi}_{\mathcal{I}}^2 + \epsilon
\end{aligned} \tag{14}$$

where the implication (e) follows from the fact that the matrix $I - \mathcal{O}_{\mathcal{I}}\mathcal{O}_{\mathcal{I}}^{+}$ is idempotent, hence $\left\|I - \mathcal{O}_{\mathcal{I}}\mathcal{O}_{\mathcal{I}}^{+}\right\|_2^2 \leq 1$; (f) follows from the properties of the induced 2-norm, which implies that for any matrix A and vector z then $\|Az\|_2 \leq \|A\|_2 \|z\|_2$, and hence $\left\|(I - \mathcal{O}_{\mathcal{I}}\mathcal{O}_{\mathcal{I}}^{+})\Psi_{\mathcal{I}}\right\|_2 \leq \left\|I - \mathcal{O}_{\mathcal{I}}\mathcal{O}_{\mathcal{I}}^{+}\right\|_2 \|\Psi_{\mathcal{I}}\|_2$; (g) follows from Proposition IV.2. Finally, (h) follows from the fact that $(\overline{\Psi}_{\mathcal{I}}^2 + \sqrt{\epsilon})^2 \geq \overline{\Psi}_{\mathcal{I}}^2 + \epsilon$.

Combining the bounds (13) and (14), we conclude that the following holds

$$\|Y_{\mathcal{I}} - \mathcal{O}_{\mathcal{I}} x\|_2^2 > \overline{\Psi}_{\mathcal{I}}^2 + \epsilon$$

which implies that the result of Algorithm 2 is UNSAT whenever (12) is satisfied.

The error bound δ can be then computed directly as

$$\begin{aligned}
\|x^* - x\|_2^2 &= \left\|x^* - \mathcal{O}_{\mathcal{I}}^{+} Y_{\mathcal{I}}\right\|_2^2 \overset{(i)}{=} \left\|\mathcal{O}_{\mathcal{I}}^{+} \Psi_{\mathcal{I}}\right\|_2^2 \\
&\leq \left\|\mathcal{O}_{\mathcal{I}}^{+}\right\|_2^2 \|\Psi_{\mathcal{I}}\|_2^2 \overset{(j)}{\leq} \overline{o}\overline{\Psi}^2
\end{aligned}$$

where the equality (i) follows from the fact that all attacks satisfy (12) and hence can be detected. Accordingly, the set $\mathcal{I}$ contains only sensors that are attack free and therefore (3) can be simplified into $Y_{\mathcal{I}} = \mathcal{O}_{\mathcal{I}} x^* + \Psi_{\mathcal{I}}$, which in return implies that

$$\mathcal{O}_{\mathcal{I}}^{+}\, Y_{\mathcal{I}} = \mathcal{O}_{\mathcal{I}}^{+}\, \mathcal{O}_{\mathcal{I}} x^* + \mathcal{O}_{\mathcal{I}}^{+}\, \Psi_{\mathcal{I}} = x^* + \mathcal{O}_{\mathcal{I}}^{+}\, \Psi_{\mathcal{I}}.$$

Finally, inequality (j) follows from the definition of $\overline{o}$ in Definition IV.1. ■

Remark IV.4: The proof of Theorem IV.3 only relies on following assumption:

$$\|E_{\mathcal{I}}\|_2 > \left(\frac{2}{\sqrt{1-\Delta_s}}\right) \overline{\Psi}_{\mathcal{I}} + \frac{\sqrt{\epsilon}}{\sqrt{1-\Delta_s}}.$$

However, the set $\mathcal{I}$ is not known *a priori*, since it will be selected by the SAT solver; we then need to resort to the more conservative assumption in the statement of Theorem IV.3:

$$\|E_i\|_2 > \left(\frac{2}{\sqrt{1-\Delta_s}}\right) \overline{\Psi} + \frac{\sqrt{\epsilon}}{\sqrt{1-\Delta_s}}$$

which will also be used in Theorem IV.5 below.

Theorem IV.3 characterizes the class of attack signals that lead to detection. However, a smart attacker may be tempted to inject attack signals that are not detected by the proposed algorithm, yet increase the estimation error. The following result characterizes the estimation error in the presence of un-detectable attacks.

Theorem IV.5: Let the linear system defined in (1) be $2\overline{s}$-sparse observable, let $\epsilon > 0$ be the numerical solver tolerance. Then, Algorithm 1, modified as in (8), returns an estimate x that satisfies:

$$\|x^* - x\|_2 \leq \overline{o}\left(1 + \frac{2}{\sqrt{1-\Delta_s}}\right) \overline{\Psi} + \frac{\overline{o}\sqrt{\epsilon}}{\sqrt{1-\Delta_s}}.$$

Proof: The error $\|x^* - x\|_2$ can be bounded as follows:

$$\begin{aligned}
\|x^* - x\|_2 &= \left\|x^* - \mathcal{O}_{\mathcal{I}}^{+}\, Y_{\mathcal{I}}\right\|_2 \\
&= \left\|x^* - \mathcal{O}_{\mathcal{I}}^{+}\, \mathcal{O}_{\mathcal{I}} x^* - \mathcal{O}_{\mathcal{I}}^{+}\, \Psi_{\mathcal{I}} - \mathcal{O}_{\mathcal{I}}^{+}\, E_{\mathcal{I}}\right\|_2 \\
&= \left\|\mathcal{O}_{\mathcal{I}}^{+}\, \Psi_{\mathcal{I}} + \mathcal{O}_{\mathcal{I}}^{+}\, E_{\mathcal{I}}\right\|_2 \\
&\overset{(a)}{\leq} \left\|\mathcal{O}_{\mathcal{I}}^{+}\right\|_2 \|\Psi_{\mathcal{I}}\|_2 + \left\|\mathcal{O}_{\mathcal{I}}^{+}\right\|_2 \|E_{\mathcal{I}}\|_2 \\
&\overset{(b)}{\leq} \overline{o}\overline{\Psi} + \overline{o}\, \|E_{\mathcal{I}}\|_2 \\
&\overset{(c)}{\leq} \overline{o}\overline{\Psi} + \overline{o}\frac{2}{\sqrt{1-\Delta_s}}\, \|\Psi\|_2 + \overline{o}\frac{\sqrt{\epsilon}}{\sqrt{1-\Delta_s}} \\
&= \overline{o}\left(1 + \frac{2}{\sqrt{1-\Delta_s}}\right) \overline{\Psi} + \frac{\overline{o}\sqrt{\epsilon}}{\sqrt{1-\Delta_s}}
\end{aligned}$$

where inequality (a) follows from Cauchy–Schwarz inequality; (b) follows from the definition of $\overline{o}$ in IV.1 along with the fact that $\|\Psi_{\mathcal{I}}\|_2 \leq \|\Psi\|_2$; (c) follows from Theorem IV.3 (along with Remark IV.4), stating that only attacks with norm

$$\|E_{\mathcal{I}}\|_2 \leq \left(\frac{2}{\sqrt{1-\Delta_s}}\right) \overline{\Psi} + \frac{\sqrt{\epsilon}}{\sqrt{1-\Delta_s}}$$

may not be detected by Algorithm 1 and hence can affect the estimation error. ■

V. Experimental Results

We developed our theory solver in MATLAB, and interfaced it with the pseudo-Boolean SAT solver SAT4J [31]. All the experiments were executed on an Intel Core i7 3.4-GHz processor with 8 GB of memory. To validate our approach, we first compare the effect of the two proposed certificates on the required number of iterations. We then compare the runtime performance against previously proposed algorithms. Finally, we demonstrate the effect of attack detection on the problem of controlling a robotic vehicle under sensor attacks. The source code for the solver and all the experiments in this section is publicly available at [32].

A. Runtime Performance

To assess the effectiveness of the algorithms introduced in Sections III-D and III-E, Fig. 3(a) shows the number of iterations of IMHOTEP-SMT when only one of the three certificates, the trivial certificate $\phi_{\text{triv-cert}}$, the conflicting certificate $\phi_{\text{conf-cert}}$, and the joint certificate $\phi_{\text{conf-cert}} \wedge \phi_{\text{agree-cert}}$, is used. In each test case, we generated a random support set for the attack vector, a random attack signal, and random initial conditions. All reported results are averaged results over 20 runs of the same experiment. Although we claim no statistical significance, the results reported in this section are representative of the several simulations performed by the authors.

In the first experiment (top), we increase the number s of actual sensors under attack for a fixed $\overline{s} = 20$ $(n = 25, p = 60)$. In the second experiment (bottom), we increase both n and p simultaneously, with $p = 3n$, while $p/3$ sensors are under attack, and $\overline{s} = p/3$. In both cases, the system is constructed to be $3\overline{s}$-sparse observable, with the dimensions of the kernels of $\mathcal{O}_i$ ranging between $n-1$ and $n-2$, meaning that the state is "poorly" observable from individual sensors. We also show the number of iterations against the theoretical limit in Proposition III.5. We observed an average of 50× reduction in iterations when $\phi_{\text{conf-cert}}$ was used compared to $\phi_{\text{triv-cert}}$, while using both $\phi_{\text{conf-cert}}$ and $\phi_{\text{agree-cert}}$ decreased the number of iterations by a factor of 75.

We also compared the performance of IMHOTEP-SMT against the MIQP formulation (5), the event-triggered projected gradient descent (ETPG) algorithm [16], and the l_1/l_r decoder [7], with respect to both execution time and estimation error. The MIQP is solved using the commercial solver GUROBI [33], the ETPG algorithm is implemented in MATLAB, while the l_1/l_r decoder is implemented using the convex solver CVX [34].

Fig. 3 reports the numerical results in two test cases. In Fig. 3(b), we fix the number of sensors $p = 20$ and increase the number of system states from $n = 10$ to $n = 150$. In Fig. 3(c), we fix the number of states $n = 50$ and increase the number of sensors from $p = 3$ to $p = 150$. In both cases, half of the sensors are attacked. Our algorithm always outperforms both the ETPG and the l_1/l_r approaches and scales nicely with respect to both n and p. In particular, as evident from Fig. 3(b), increasing n has a small effect on the overall execution time, which reflects the fact that the number of constraints to be satisfied does not depend on n. Conversely, as shown in Fig. 3(c), as the number of sensors increases, the number of constraints, hence the execution time of our algorithm, also increases. The runtime of the MIQP formulation in (5) scales worse than our algorithm

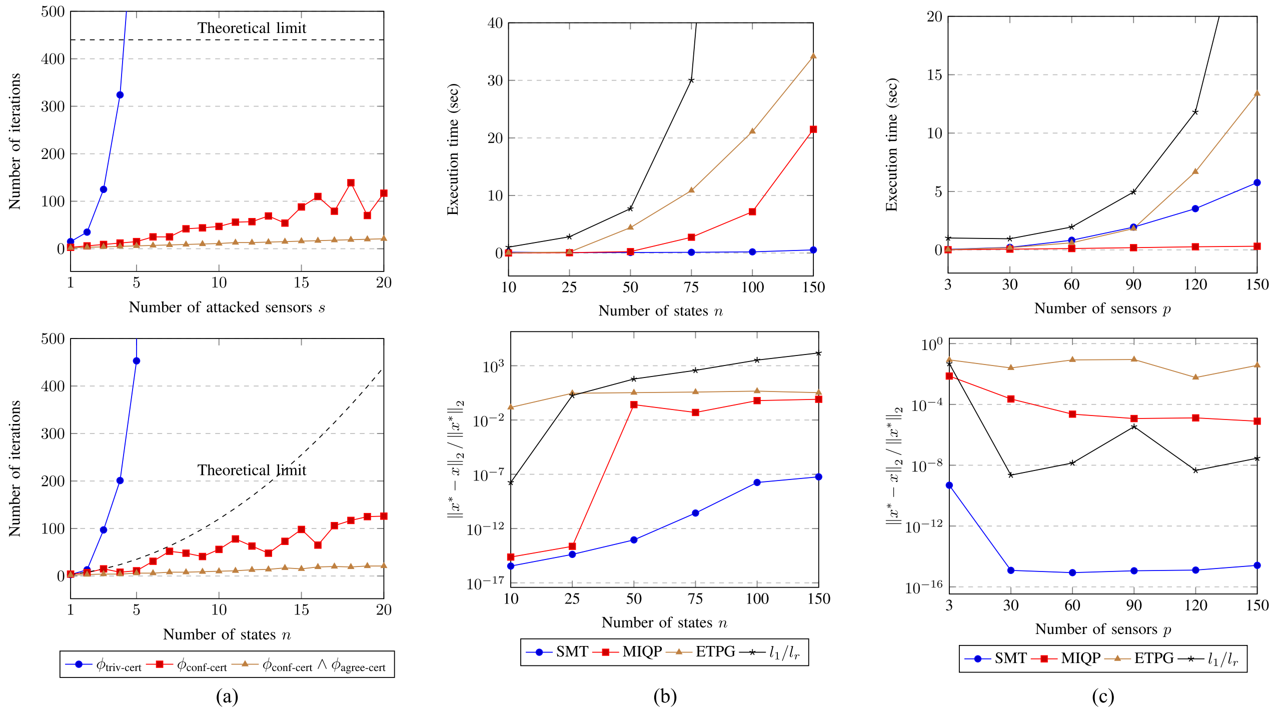


Fig. 3. Simulation results showing number of iterations, execution time, and estimation error with respect to number of states and number of sensors. (a) Number of iterations in Algorithm 1 versus number of attacked sensors (top) and number of states and sensors (down) for different strategies to generate compact certificates. (b) Execution time (top) and estimation error (bottom) versus number of states n for different algorithms ($p = 20$, $\overline{s} = 5$). (c) Execution time (top) and estimation error (bottom) versus number of sensors p for different algorithms ($n = 50$, $\overline{s} = p/2 - 1$).

with n, but better with p, because GUROBI can efficiently process many conic constraints (whose number scales with p) but is more sensitive to the size of each conic constraint (which scales with n). Finally, Fig. 3(b) (bottom) shows that the l_1/l_r decoder reports incorrect results in multiple test cases, because of its lack of soundness, as anticipated in Section I.

B. Securing an UGV

We apply our algorithms to the model of a UGV, as detailed in [12] and [16], under different types of sensor attacks. We assume that the UGV moves along straight lines and completely stops before rotating. Under these assumptions, we can describe the dynamics of the UGV as

$$\begin{bmatrix} \dot{x} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & \frac{-B}{M} \end{bmatrix} \begin{bmatrix} x \\ v \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{M} \end{bmatrix} F$$

where x and v are the states, corresponding to the UGV position and linear velocity, respectively. The parameters M and B denote the mechanical mass and the translational friction coefficient. The inputs to the UGV is the force F. The UGV is equipped with a GPS sensor which measures its position and two motor encoders which measure the translational velocity. The resulting output equation is

$$y = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ v \end{bmatrix} + \begin{bmatrix} \psi_1 \\ \psi_2 \\ \psi_3 \end{bmatrix}$$

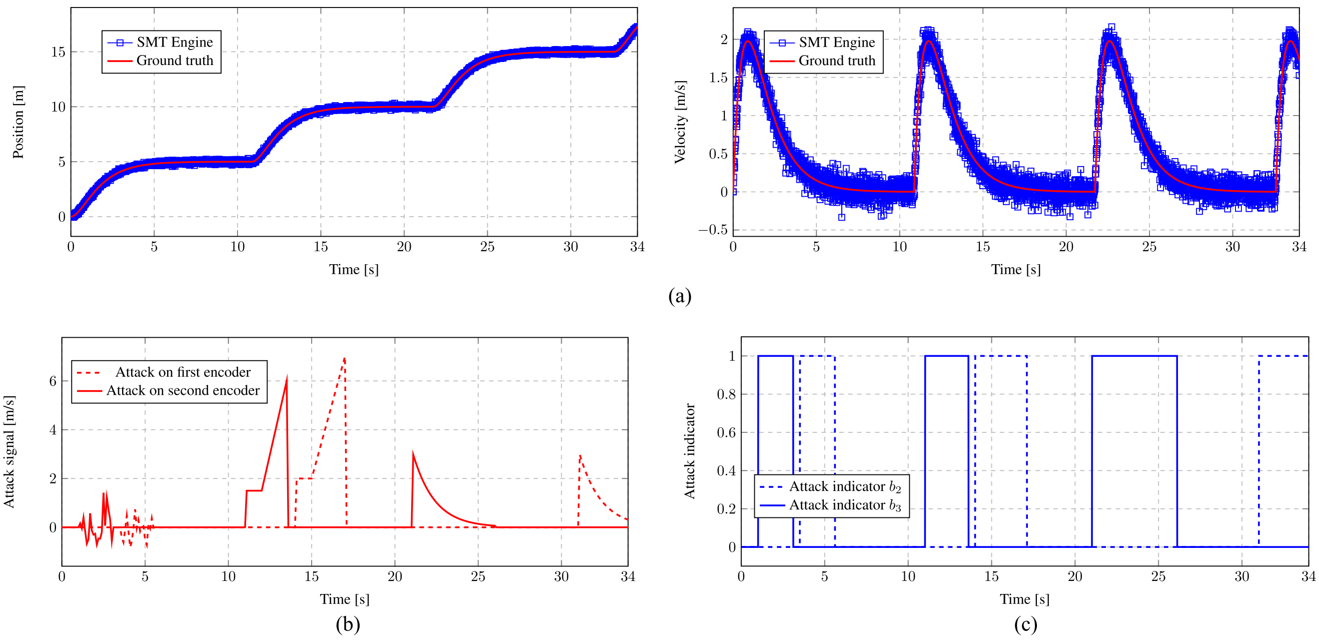
where ψ_i is the measurement noise on the ith sensor that is assumed to be bounded. In our experiments, we used $M = 0.8$ kg, $B = 1$, $\overline{\psi}_1^2 = 0.2$ m^2, $\overline{\psi}_2^2 = \overline{\psi}_3^2 = 0.2$ (m/s)^2.

The model is discretized with a time step equal to 0.1 s. The SMT-based detector uses the discretized model along with sensor measurements to provide an estimate for the state vector, which is then used by a feedback controller to regulate the robot and follow a squared-shape path of length equal to 5 m.

Fig. 4 shows the performance of the SMT-based detector. The attacker alternates between corrupting the first and the second encoder measurements as shown in Fig. 4(b). Three different types of attacks are considered. First, the attacker corrupts the sensor signal with random noise. The next attack consists of a step function followed by a ramp. Finally, a replay attack is mounted by replaying the previously measured UGV velocity. The estimated position and velocity are shown in Fig. 4(a). We recall that the SMT-based detector is also able to return the indicator variable vector b, denoting which sensors are under attack. Fig. 4(b) shows both the attack and the corresponding indicator variables as returned by the SMT-based detector. The proposed algorithm is able to estimate the state and the support of the attack also in the presence of noise.

VI. Conclusion

We proposed a sound and complete algorithm that adopts a SMT paradigm to tackle the intrinsic combinatorial complexity of the secure state estimation problem for linear dynamical systems under sensor attacks and in the presence of noise. At the heart of our detector lies a set of routines that exploit the geometric structure of the problem to efficiently reason about inconsistency of sensor measurements and enhance the runtime performance. Our approach was validated via numerical simulations and demonstrated on an UGV control problem. Future



(a)

(b)

(c)

Fig. 4. Performance of the UGV controller in the case when no attack takes place versus the case when the attack signal is applied to the UGV encoders. The objective is to move 5 m, stop and perform a 90° rotation, and repeat this pattern to follow a square path. The controller uses the proposed SMT-based approach to estimate the UGV states. In both cases, we show the linear position and linear velocity (top), and the attack signal and its estimate (bottom). (a) Estimated position and velocity versus ground truth. (b) Attack signal on the two encoders. (c) Indicator variables b computed by the proposed SMT-based detector.

directions include the extension and the characterization of the proposed algorithm for nonlinear and hybrid dynamical systems.

Appendix

In Proposition A.1, we recall a general result that will be used in the proof of Proposition IV.2. In order to state this result, we first recall the following two facts.

Fact 1: For any two square matrices A and B, both AB and BA have the same eigenvalues.

Fact 2: If $I - A$ is a positive definite matrix, then all the eigenvalues of A are strictly less than 1.

Proposition VI.1: Given a positive semidefinite matrix A and a positive definite matrix B of the same dimension, then every eigenvalue of $A(A+B)^{-1}$ is strictly less than 1.

Proof: It follows from the positive (semi)definiteness assumptions of A and B that $(A+B)^{-1}$ is positive definite matrix, hence it can be written using its square root matrix as

$$(A+B)^{-1} = (A+B)^{-\frac{1}{2}}(A+B)^{-\frac{1}{2}}.$$

It also follows from Fact 1 that $A(A+B)^{-1}$ has the same eigenvalues of $(A+B)^{-\frac{1}{2}}A(A+B)^{-\frac{1}{2}}$. Therefore, we obtain

$$\begin{aligned} I - (A+B)^{-\frac{1}{2}}A(A+B)^{-\frac{1}{2}} \\ &= (A+B)^{-\frac{1}{2}}(A+B)(A+B)^{-\frac{1}{2}} \\ &- (A+B)^{-\frac{1}{2}}A(A+B)^{-\frac{1}{2}} \\ &= (A+B)^{-\frac{1}{2}}B(A+B)^{-\frac{1}{2}} \end{aligned}$$

which is still positive definite. Hence, from Fact 2, all eigenvalues of $(A+B)^{-\frac{1}{2}}A(A+B)^{-\frac{1}{2}}$ are strictly less than 1, which implies that also the eigenvalues of $A(A+B)^{-1}$ are strictly less than 1. ■

Acknowledgment

The US Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF, DARPA, or the US Government.

References

[1] R. Langner, "Stuxnet: Dissecting a cyberwarfare weapon," *IEEE Security Privacy*, vol. 9, no. 3, pp. 49–51, May 2011.

[2] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," in *Proc. ACM Conf. Comput. Commun. Security*, New York, NY, USA, Nov. 2009, pp. 21–32.

[3] Y. Shoukry, P. D. Martin, P. Tabuada, and M. B. Srivastava, "Non-invasive spoofing attacks for anti-lock braking systems," in *Proc. Workshop Cryptographic Hardw. Embedded Syst.*, 2013, pp. 55–72.

[4] C.-Z. Bai and V. Gupta, "On Kalman filtering in the presence of a compromised sensor: Fundamental performance bounds," in *Proc. Amer. Control Conf.*, Jun. 2014, pp. 3029–3034.

[5] Y. Mo and B. Sinopoli, "Secure estimation in the presence of integrity attacks," *IEEE Trans. Autom. Control*, vol. 60, no. 4, pp. 1145–1151, Apr. 2015.

[6] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *Proc. Allerton Conf. Commun., Control, Comput.*, Sep. 2009, pp. 911–918.

[7] 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.

[8] F. Pasqualetti, F. Dorfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Trans. Autom. Control*, vol. 58, no. 11, pp. 2715–2729, Nov. 2013.

[9] M. S. Chong, M. Wakaiki, and J. P. Hespanha, "Observability of linear systems under adversarial attacks," in *Proc. Amer. Control Conf.*, Jul. 2015, pp. 2439–2444.

[10] C. Lee, H. Shim, and Y. Eun, "Secure and robust state estimation under sensor attacks, measurement noises, and process disturbances: Observer-based combinatorial approach," in *Proc. Eur. Control Conf.*, Jul. 2015, pp. 1872–1877.

[11] 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, pp. 2929–2933.

[12] M. Pajic, J. Weimer, N. Bezzo, P. Tabuada, O. Sokolsky, I. Lee, and G. J. Pappas, "Robustness of attack-resilient state estimators," in *Proc. ACM/IEEE Int. Conf. Cyber-Physical Syst.*, Apr. 2014, pp. 163–174.

[13] Y. H. Chang, Q. Hu, and C. J. Tomlin, "Secure estimation based Kalman filter for cyber-physical systems against adversarial attacks," arXiv:1512.03853, Jun. 2015.

[14] S. Yong, M. Zhu, and E. Frazzoli, "Resilient state estimation against switching attacks on stochastic cyber-physical systems," in *Proc. IEEE Int. Conf. Decision Control*, Dec. 2015, pp. 5162–5169.

[15] Y. Shoukry and P. Tabuada, "Event-triggered projected Luenberger observer for linear systems under sparse sensor attacks," in *Proc. IEEE Int. Conf. Decision Control*, Dec. 2014, pp. 3548–3553.

[16] Y. Shoukry and P. Tabuada, "Event-triggered state observers for sparse sensor noise/attacks," *IEEE Trans. Autom. Control*, vol. 61, no. 8, pp. 2079–2091, Aug. 2016.

[17] A. Tiwari *et al.*, "Safety envelope for security," in *Proc. Int. Conf. High Confidence Networked Syst.*, Apr. 2014, pp. 85–94.

[18] 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.

[19] 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.

[20] M. Blanke, M. Kinnaert, J. Lunze, M. Staroswiecki, and J. Schröder, *Diagnosis and Fault-Tolerant Control*, vol. 2. New York, NY, USA: Springer, 2006.

[21] C. Barrett, R. Sebastiani, S. A. Seshia, and C. Tinelli, "Satisfiability modulo theories," in *Handbook of Satisfiability*. Amsterdam, The Netherlands: IOS Press, 2009.

[22] P. Nuzzo, A. Puggelli, S. A. Seshia, and A. Sangiovanni-Vincentelli, "CalCS: SMT solving for non-linear convex constraints," in *Proc. Formal Methods Comput.-Aided Des.*, Oct. 2010, pp. 71–79.

[23] Y. Shoukry, A. Puggelli, P. Nuzzo, A. L. Sangiovanni-Vincentelli, S. A. Seshia, and P. Tabuada, "Sound and complete state estimation for linear dynamical systems under sensor attack using satisfiability modulo theory solving," in *Proc. Amer. Control Conf.*, Jul. 2015, pp. 3818–3823.

[24] Y. Shoukry *et al.*, "Imhotep-SMT: A satisfiability modulo theory solver for secure state estimation," in *Proc. Int. Workshop Satisfiability Modulo Theories*, Jul. 2015, pp. 3–13.

[25] C. W. Brown and J. H. Davenport, "The complexity of quantifier elimination and cylindrical algebraic decomposition," in *Proc. Int. Symp. Symbolic Algebraic Comput.*, 2007, pp. 54–60.

[26] G. E. Collins, "Quantifier elimination for real closed fields by cylindrical algebraic decomposition: A synopsis," *SIGSAM Bull.*, vol. 10, no. 1, pp. 10–12, Feb. 1976.

[27] S. Gao, J. Avigad, and E. M. Clarke, "δ-complete decision procedures for satisfiability over the reals," in *Proc. 6th Int. Joint Conf. Autom. Reasoning*, Jun. 2012, pp. 286–300.

[28] S. Sundaram and C. N. Hadjicostis, "Distributed function calculation via linear iterative strategies in the presence of malicious agents," *IEEE Trans. Autom. Control*, vol. 56, no. 7, pp. 1495–1508, Jul. 2011.

[29] W. L. Winston, *Operations Research: Applications & Algorithms*. Stamford, Connecticut, US: Thomson Business Press, 2008.

[30] R. Nieuwenhuis, A. Oliveras, and C. Tinelli, "Solving SAT and SAT modulo theories: From an abstract Davis–Putnam–Logemann–Loveland procedure to DPLL(T)," *J. ACM*, vol. 53, no. 6, pp. 937–977, Nov. 2006.

[31] D. L. Berre and A. Parrain, "The Sat4j library, release 2.2," *J. Satisfiability, Boolean Modeling Comput.*, vol. 7, pp. 59–64, 2010.

[32] "Imhotep-SMT," Github repository, Jan. 2015. [Online]. Available: http://nesl.github.io/Imhotep-smt/.

[33] Gurobi Optimizer, Jan. 2015. [Online]. Available: http://www.gurobi.com/.

[34] M. Grant and S. Boyd, "CVX: MATLAB software for disciplined convex programming, version 1.21," May 2010. [Online]. Available: http://cvxr.com/cvx.

Yasser Shoukry received the B.Sc. and M.Sc. degrees (with distinction and honors) in computer and systems engineering from Ain Shams University, Cairo, Egypt, in 2010 and 2007, respectively, and the Ph.D. degree in electrical engineering from the University of California at Los Angeles, Los Angeles, CA, USA, in 2015.

He was affiliated with both the Cyber-Physical Systems Laboratory as well as the Networked and Embedded Systems Laboratory, University of California at Los Angeles, Los Angeles, CA, USA. He is a Postdoctoral Scholar at both the Department of Electrical Engineering and Computer Sciences of the University of California, Berkeley, CA, USA and the Department of Electrical Engineering, University of California, Los Angeles, CA, USA. Before joining UCLA, he spent four years in the industry of automotive embedded systems where he was an R&D Engineer in the domain of optimizing and automatic testing of embedded software and model-based designs. His research interests include the design and implementation of secure- and privacy-aware cyber-physical systems by drawing on tools from control theory, optimization theory, embedded systems, and formal methods.

Dr. Shoukry received the Best Paper Award from the International Conference on Cyber-Physical Systems (ICCPS) in 2016. He is also received the UCLA EE Distinguished Ph.D. Dissertation Award in 2016, the UCLA Chancellors prize in 2011 and 2012, the UCLA EE Graduate Division Fellowship in 2011 and 2012, and the UCLA EE Preliminary Exam Fellowship in 2012.

Pierluigi Nuzzo (S'06–M'16) received the Laurea degree in electrical engineering (summa cum laude) from the University of Pisa, Pisa, Italy, in 2003, the Diploma degree in engineering (summa cum laude) from the Sant'Anna School of Advanced Studies, Pisa, Italy, in 2004, the Ph.D. degree in electrical engineering and computer sciences from the University of California at Berkeley, Berkeley, CA, USA, in 2015.

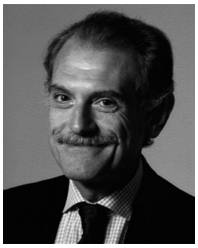
Before joining the University of California at Berkeley, he was a Researcher at IMEC, Leuven, Belgium, working on the design of energy-efficient A/D converters and frequency synthesizers for reconfigurable radio. From 2004 to 2006, he was with the Department of Information Engineering, University of Pisa, and with IMEC, as a Visiting Scholar, working on low power A/D converter design for wide-band communications and design methodologies for mixed-signal integrated circuits. He joined the Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA, in 2016, as an Assistant Professor. His current research interests include: methodologies and tools for cyber-physical system and mixed-signal system design; contracts, interfaces, and compositional methods for embedded system design; the application of formal methods and optimization theory to problems in embedded and cyber-physical systems and electronic design automation.

Prof. Nuzzo received First Place in the operational category and Best Overall Submission in the 2006 DAC/ISSCC Design Competition, a Marie Curie Fellowship from the European Union in 2006, the University of California at Berkeley EECS departmental fellowship in 2008, the University of California at Berkeley Outstanding Graduate Student Instructor Award in 2013, the IBM Ph.D. Fellowship in 2012 and 2014, the Best Paper Award from the International Conference on Cyber-Physical Systems in 2016, and the David J. Sakrison Memorial Prize in 2016.

Alberto Puggelli (S'09) received the B.Sc. and two M.Sc. degrees in electrical engineering (summa cum laude) from Politecnico di Milano, Milan, Italy, and Politecnico di Torino, Tourin, Italy, in 2006 and 2008, respectively. He received the M.Sc. degree in computer science and the Ph.D. degree in electrical engineering and computer science from the University of California at Berkeley, Berkeley, CA, USA in 2013 and 2014, respectively.

He was with ST-Ericsson in 2009 and with Texas Instruments in 2011 and 2012, as an Intern Analog Designer. He is currently the Director of Technology at Lion Semiconductor Inc., San Francisco, CA, USA. He is the author or coauthor of more than 20 publications in IEEE/ACM conference proceedings and journals. He holds four US patents. His research interests include the design of hybrid dc–dc voltage regulators.

Dr. Puggelli received the two Gold Medal Awards for the Best Student from the Politecnico di Milano. He received the AEIT Fellowship Isabella Sassi Bonadonna in 2010.



Alberto L. Sangiovanni-Vincentelli received Doctor of Engineering, Electrical Engineering and Computer Science from Politecnico di Milano, Italy. He has also two honorary EECS doctorates from Aalborg University, Denmark, and KTH, Stockholm, Sweden. He holds the Buttner Chair of electrical engineering and computer sciences, University of California, Berkeley, CA, USA. He is on the Advisory Board of three companies and has consulted for companies such as Intel, HP, Bell Labs, IBM, Samsung, UTC, Kawasaki Steel, Fujitsu, Telecom Italia, Pirelli, BMW, Mercedes, Magneti Marelli, ST Microelectronics, LElettronica, and UniCredit. He is an author of more than 850 papers, 17 books, and 2 patents.

Mr. Sangiovanni-Vincentelli for his scientific research, received the IEEE/RSE James Clerk Maxwell Award for “groundbreaking contributions that have had an exceptional impact on the development of electronics and electrical engineering or related fields,” the Kaufmann Award for seminal contributions to EDA, the IEEE Darlington Award, the IEEE Guillemin-Cauer Award, the EDAA lifetime Achievement Award, the IEEE/ACM R. Newton Impact Award, the University of California Distinguished Teaching Award, the SRC Aristotle Award, and the IEEE Graduate Teaching Award for inspirational teaching of graduate students. He is a member of the National Academy of Engineering, and holds two honorary Doctorates. On the industrial side, he helped founding Cadence and Synopsys, the two leading companies in Electronic Design Automation and is on the Board of five companies including Cadence.

Sanjit A. Seshia (S'99–M'05–SM'11) received the B.Tech. degree in computer science and engineering from the Indian Institute of Technology, Bombay, India, in 1998, and the M.S. and Ph.D. degrees in computer science from Carnegie Mellon University, Pittsburgh, PA, USA, in 2000 and 2005, respectively.

He is currently an Associate Professor in the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA. His research interests include dependable computing and computational logic, with a current focus on applying automated formal methods to problems in embedded and cyber-physical systems, electronic design automation, computer security, and synthetic biology. His Ph.D. thesis work on the UCLID verifier and decision procedure helped pioneer the area of satisfiability modulo theories (SMT) and SMT-based verification. He is the coauthor of a widely used textbook on embedded systems. He led the offering of a massive open online course on cyber-physical systems for which his group developed novel virtual lab auto-grading technology based on formal methods.

Prof. Seshia has served as an Associate Editor of the IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS, and as a Co-Chair of the Program Committee of the International Conference on Computer-Aided Verification in 2012. His awards and honors include the Presidential Early Career Award for Scientists and Engineers from the White House, the Alfred P. Sloan Research Fellowship, the Prof. R. Narasimhan Lecture Award, and the School of Computer Science Distinguished Dissertation Award at Carnegie Mellon University.

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

Between January 2002 and July 2003, he was a Postdoctoral researcher at the University of Pennsylvania. After spending three years at the University of Notre Dame, as an Assistant Professor, he joined the Electrical Engineering Department, University of California, Los Angeles, CA, USA, where he established and directs the Cyber-Physical Systems Laboratory. His latest book, *Verification and Control of Hybrid Systems* (Springer, 2009).

Dr. Tabuada's contributions to cyber-physical systems have been recognized by multiple awards including the NSF 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 on Hybrid Systems: Computation and Control and joined its steering committee in 2015; in 2012, he was a Program Co-Chair for the 3rd IFAC Workshop on Distributed Estimation and Control in Networked Systems, and in 2015, he was a 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.