# Observability of linear systems under adversarial attacks*

Michelle S. Chong[1], Masashi Wakaiki[2] and João P. Hespanha[2]

***Abstract*— We address the problem of state estimation for multi-output continuous-time linear systems, for which an attacker may have control over some of the sensors and inject (potentially unbounded) additive noise into some of the measured outputs. To characterize the resilience of a system against such sensor attacks, we introduce a new notion of observability — termed "observability under attacks" — that addresses the question of whether or not it is possible to uniquely reconstruct the state of the system by observing its inputs and outputs over a period of time, with the understanding that some of the available system's outputs may have been corrupted by the opponent. We provide computationally efficient tests for observability under attacks that amount to testing the (standard) observability for an appropriate finite set of systems. In addition, we propose two state estimation algorithms that permit the state reconstruction in spite of the attacks. One of these algorithms uses observability Gramians and a finite window of measurements to reconstruct the initial state. The second algorithm takes the form of a switched observer that asymptotically converges to the correct state estimate in the absence of additive noise and disturbances, or to a neighborhood of the correct state estimate in the presence of bounded noise and disturbances.**

## I. Introduction

This paper is motivated by the observation that computer control systems can be especially vulnerable to cyber attacks; most particularly remote sensors that can be infiltrated and reprogrammed to report erroneous measurements.

The issue of security is not new to the control field, in particular in the areas of fault detection and identification (FDI) [8] and game theory [5], [7]. Some of the recent work on the cyber security of control systems have been focused on the effect of specific types of attacks on stability and/or estimation [14], such as false data injection attacks [6], [3], denial-of-service attacks [1], [13] and integrity attacks [9]. Closer to the work presented here, there has also been an effort to derive results that are independent of the attack type in works such as [2], [11] and [10]. In [10], the authors model the attacked system as a continuous-time descriptor system and view the attack signal as an unknown input. The authors then propose an algorithm that detects the presence of an attack. Other related works focused on robust state estimation appeared in [2], [11], where the authors characterized the resilience of a discrete-time LTI system against attacks by the number of attacked sensors allowed for accurate state reconstruction. They also proposed an error correction algorithm which exactly reconstructs the state and is made computationally efficient by transforming the optimization problem into a convex one, which is possible only under certain conditions. Due to the close relation of these works to our paper which we were unaware of at the time of writing, we provide a comparison after we have outlined our results.

The scenario considered in this paper considers a continuous-time LTI system with $N$ outputs, each measured by a potentially vulnerable sensor. One then asks whether or not it is possible to reconstruct the initial state of the system from an input/output time series if $M \leqslant N$ of these sensors have been taken over by an adversary. It is assumed that the attacker has full control over the measurement reported by the $M$ infiltrated sensors, with the understanding that we do not know which of the $M$ sensors have been infiltrated and, in fact, if they have been infiltrated at all. When it is possible to do state reconstruction under this scenario, we say that the LTI system is *observable under $M$ attacks*.

The first key result of this paper is a necessary and sufficient condition presented in Section II for observability under $M$ attacks. This condition requires the number of sensors $N$ to be larger than $2M$ and also that a family of LTI systems (derived from our original system) is observable, under the usual notion of observability. It was expected for $N > M$ to be necessary for an $N$-output system to be observable under $M$ attacks since with $N \leqslant M$ there would be no attack-free measurements left to use for estimation. However, it is somewhat unexpected to see that $N > 2M$ is actually necessary for an $N$-output system to be observable under $M$ attacks.

The second key result of the paper is an algorithm presented in Section III-A that looks at the values of the system's input and all $N$ measured outputs over a finite interval $[0, T]$, $T < \infty$ and provides a correct estimate of the system's initial state, in spite of the fact that $M$ of the $N$ measured outputs may have been compromised by an attacker. As expected, this algorithm is only applicable to systems that are observable under $M$ attacks. In essence, the algorithm proposed constructs multiple state estimates using observability Gramians and utilizes a consistency condition to select the "correct" estimate.

This material is based upon work supported by the National Science Foundation under Grant No. CNS-1329650 and by the U.S. Army Research Laboratory and the U.S. Army Research Office under MURI grant No. W911NF-09-1-0553. M. Chong acknowledges the American Australian Association for their support of this work. M. Wakaiki is supported by The Kyoto University Foundation.

[1]The author is with the Department of Electrical and Electronic Engineering, the University of Melbourne, Australia. This work was conducted when the author was at the Center for Control, Dynamical-systems and Computation (CCDC), University of California, Santa Barbara, CA 93106-9560 USA. mstchong@gmail.com

[2]The authors are with the Center for Control, Dynamical-systems and Computation (CCDC), University of California, Santa Barbara, CA 93106-9560 USA. {masashiwakaiki, hespanha}@ece.ucsb.edu

The third key result is an observer-like algorithm presented in Section III-B that (causally) creates an asymptotically correct estimate of the system's current state based on the values of its past input and $N$ (potentially compromised) outputs. This algorithm is applicable when the system is observable under $M$ attacks, but it actually requires less than that. In practice, it only requires a notion of detectability under $M$ attacks. This is not surprising in view of the fact that this observer-like algorithm does not "promise" state reconstruction in finite time (only asymptotically). For this algorithm, we actually prove more than just asymptotically correct state reconstruction, as we also show that an additive bounded disturbance and additive bounded measurement noise to all $N$ outputs will result in a bounded estimation error, by providing an input-to-state stability-like bound on the estimation error, in terms of bounds on the disturbance and measurement noise [12].

*Relation to similar works [2], [11]*

Although our results differ from the works [2], [11] where the discrete-time setting is considered, we address similar problems in continuous-time. Therefore, we provide a comparison:

- The authors of [2], [11] provide necessary and sufficient conditions which involve combinatorially checking the observability of a family of LTI discrete-time system. Our results in continuous-time are consistent with the results reported in the aforementioned works. The difference lies in [2] where the successful reconstruction of the states is formulated as an optimization problem, whereas we use a form of observability where the initial state of the system can be reconstructed from known inputs and outputs in the presence of attacks.
- In [2], [11], algorithms are proposed to estimate the state and attack signals from available measurements, where the $L_r\backslash L_1$ optimization-based algorithms in [2] are made more computationally efficient in [11]. Our proposed algorithms either utilizes the observability Gramian or the Luenberger observer in a multiple model setup, which can be computationally intensive. We are currently addressing this issue.

*Notation*

We denote the cardinality of a set $\mathcal{S}$ as $\text{card}(\mathcal{S})$. $|x|$ denotes the Euclidean norm of a vector $x \in \mathbb{R}^n$. $\|z\|_{\mathcal{T}}$ denotes the supremum norm of a signal $z$ on an interval $\mathcal{T} \subset [0, \infty)$. The binomial coefficient is denoted $\binom{a}{b}$, where $a$, $b$ are non-negative integers.

## II. Observability under attacks

Consider the following continuous-time LTI system with $N$ outputs:

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y_i &= C_i x + D_i u + \eta_i, \qquad i \in \{1, \dots, N\}, \end{aligned} \tag{1}$$

where the state vector is $x \in \mathbb{R}^{n_x}$, the input vector is $u \in \mathbb{R}^{n_u}$, the measured outputs are $y_i \in \mathbb{R}^{n_i}$, and the $\eta_i \in \mathbb{R}^{n_i}$ denote additive, possibly unbounded attack signals that cannot be measured. We denote the solution to (1) for the input $u$ and initial condition $x(0) = x_0$ as $x(t) = x(t; x_0, u)$ and the corresponding measured outputs as $y_i(t) = y_i(t; x_0, u, \eta_i)$, $\forall i \in \{1, \dots, N\}$.

We seek to derive conditions under which the initial condition $x(0)$ of (1) can be reconstructed from the measured outputs $y_i$, $\forall i$. However, we are interested in the possibility that a subset $y_i$, $i \in \mathcal{I} \subset \{1, \dots, N\}$ of the sensor outputs have been attacked, but we do not know which. Specifically, we assume that there is an unknown subset $\mathcal{I} \subset \{1, \dots, N\}$ with at most $M$ elements for which the corresponding $\eta_i$, $i \in \mathcal{I}$ are nonzero and could be unbounded. This motivates the following definition of "observability under attacks".

*Definition 1:* The system (1) is observable under $M$-attacks on the interval $[0, T]$, $T < \infty$ if for every initial conditions $x(0)$, $\bar{x}(0) \in \mathbb{R}^{n_x}$, input $u(t)$, $t \geqslant 0$, sets $\mathcal{I}_a$, $\mathcal{I}_b \subset \{1, \dots, N\}$ with at most $M$ elements, and attack vectors $\eta = (\eta_1, \dots, \eta_N) \in \mathcal{N}_{\mathcal{I}_a}$, $\bar{\eta} = (\bar{\eta}_1, \dots, \bar{\eta}_N) \in \mathcal{N}_{\mathcal{I}_b}$ we have

$$\begin{aligned} &y_i(t; x(0), u, \eta_i) = y_i(t; \bar{x}(0), u, \bar{\eta}_i), \\ &\forall t \in [0, T],\ i \in \{1, \dots, N\} \implies \qquad x(0) = \bar{x}(0). \end{aligned} \tag{2}$$

The notation $\mathcal{N}_{\mathcal{I}_a}$ denotes the set $\{(\eta_1, \dots, \eta_N) : \eta_i(t) = 0, \forall t \in [0, \infty), \forall i \notin \mathcal{I}_a\}$. □

In essence, this definition means that, when a system is observable under $M$-attacks, there is at most one initial condition that is compatible with the input signal $u$ and the measured outputs $y_i$, $i \in \{1, \dots, N\}$ on the interval $[0, T]$, regardless of which one of the $M$ sensors have been attacked and the corresponding attack signals $\eta_i$ selected by the opponent.

The following result provides a necessary and sufficient condition for system (1) to be observable under $M$-attacks, which permits checking whether a system is observable under attacks using standard observability tests [4, Section 15.9]. The proof is provided in Section II-A.

*Theorem 1:* For every integer $M \geqslant 0$, the following statements are equivalent:

(i) System (1) is observable under $M$-attacks on the time interval $[0, T]$, $T < \infty$.

(ii) $N > 2M$ and, for every set $\mathcal{J} \subset \{1, \dots, N\}$ with $\text{card}(\mathcal{J}) \geqslant N - 2M$, the pair $(A, C_{\mathcal{J}})$ is observable, where $C_{\mathcal{J}}$ is a matrix obtained by stacking all the output matrices $C_i$, $i \in \mathcal{J}$ from system (1).

□

Theorem 1 implicitly restricts the number of attacked outputs $M$ to be less than half of the number of outputs $N$, which is consistent with the result in [2] for the state estimation of discrete-time LTI systems under attacks.

*Remark 1:* Since condition (ii) in Theorem 1 does not depend on $T$, we conclude that if a system is observable under attack on the interval $[0, T_1]$, for some $T_1 < \infty$, it is also observable under attack on $[0, T_2]$, for every $T_2 < \infty$. This means that $x(0)$ can be determined from future inputs $u(t)$ and outputs $y_i(t)$, $i \in \{1, \dots, N\}$ over an arbitrarily small time interval $[0, T]$. □

*Remark 2:* By defining $M$-attack observability for a class of nonlinear systems of this form: $\dot{x} = Ax + \phi(u)$, where $\phi$ is a nonlinearity, $y_i = C_i x + \psi_i(u) + \eta_i$ for $i \in \{1, \ldots, N\}$, in the same manner as Definition 1, the results of Theorem 1 can be extended to this class of nonlinear systems under the assumption that the system is forward complete, i.e. the solution $x(t)$ exists for all $t \geqslant 0$, for any initial condition $x(0)$, input $u$ and attack signal $\eta_i$. In this case, this system is $M$-attack observable if and only if $N > 2M$ and, for every set $\mathcal{J} \subset \{1, \ldots, N\}$ with $\text{card}(\mathcal{J}) \geqslant N - 2M$, the pair $(A, C_{\mathcal{J}})$ is observable in the usual sense. □

The following simple examples illustrate the use of Theorem 1 in checking the observability of the system (1) when $M$ of the $N$ outputs are under attack.

*Example 1:* Consider the system

$$\begin{aligned} \dot{x}_1 &= x_2 + u \\ \dot{x}_2 &= a^2 x_1 - 2a x_2, \qquad a > 0, \end{aligned} \tag{3}$$

with $N = 3$ outputs

$$y_i = \left[x_1^T, x_2^T\right]^T + \eta_i, \text{ for } i \in \{1, 2, 3\}. \tag{4}$$

The system (3) with outputs (4) is observable in the usual sense. Since there are $N = 3$ outputs, the maximum allowable number of attacked outputs is $M = 1$. We will see that it is 1-attack observable by writing system (3)-(4) in the form of (1) and applying Theorem 1. There are only 3 sets $\mathcal{J}$ with $N - 2M = 1$ element: $\{1\}$, $\{2\}$, $\{3\}$ and the pairs $(A, C_1)$, $(A, C_2)$ and $(A, C_3)$ are all observable. Hence, the system is 1-attack observable. However, we will see in the next example that a system that is 1-attack observable is not necessarily 2-attack observable. □

*Example 2:* We now consider the same system (3), but with $N = 6$ outputs defined as follows

$$\begin{aligned} y_i &= x_1 + \eta_i, \quad \text{for } i \in \{1, 2, 3\}, \\ y_i &= x_2 + \eta_i, \quad \text{for } i \in \{4, 5, 6\}. \end{aligned} \tag{5}$$

First, observe that this system is observable in the usual sense. With $N = 6$, the maximum allowable number of attacked outputs is $M = 2$. However, we will see that this system is not 2-attack observable, it is only 1-attack observable.

By writing system (3) and (5) in the form of (1) and checking condition (ii) of Theorem 1, when $M = 1$, we obtain $\binom{N}{N-2M} = 15$ combinations of $\mathcal{J} = \{3, 4, 5, 6\}$, $\{2, 4, 5, 6\}$, $\{1, 4, 5, 6\}$ etc. where $\text{card}(\mathcal{J}) = 4$ ($\geqslant N - 2M$) and we need to check the observability of the pairs $\left(A, [C_3^T, C_4^T, C_5^T, C_6^T]^T\right)$, $\left(A, [C_2^T, C_4^T, C_5^T, C_6^T]^T\right)$, $\left(A, [C_1^T, C_4^T, C_5^T, C_6^T]^T\right)$, etc. Since all such pairs are observable, the system (3) with outputs defined in (5) is 1-attack observable.

However, when there are $M = 2$ attacked outputs, we obtain 15 combinations of $\mathcal{J} = \{1, 2\}$, $\{1, 3\}$, $\{1, 4\}$, $\{1, 5\}$ etc. where $\text{card}(\mathcal{J}) = 2$ ($\geqslant N - 2M$) and we see that not all pairs $(A, C_{\mathcal{J}})$ are observable, e.g. the pairs $(A, [C_1^T, C_2^T]^T)$, $(A, [C_1^T, C_3^T]^T)$, $(A, [C_2^T, C_3^T]^T)$ etc. are not observable. Therefore, this system is not 2-attack observable. □

### A. Proof of Theorem 1

We first note that, in view of the usual definition of observability for attack-free systems (e.g. [4, Definition 15.2]), condition (ii) in Theorem 1 can be equivalently re-stated as

(ii)' $N > 2M$ and for every set $\mathcal{J} \subset \{1, \ldots, N\}$ with $\text{card}(\mathcal{J}) \geqslant N - 2M$, and for every initial condition $x(0) \in \mathbb{R}^{n_x}$, we have

$$C_i e^{At} x(0) = 0, \, \forall i \in \mathcal{J}, \, t \in [0, T] \implies x(0) = 0. \tag{6}$$

We will thus prove Theorem 1 by showing that condition (i) is equivalent to (ii)' above.

**(i) $\implies$ (ii)'**: Suppose by contradiction that (i) holds, but (ii)' is false, i.e., $N \leqslant 2M$ or there exists a set $\mathcal{J} \subset \{1, \ldots, N\}$ with $\text{card}(\mathcal{J}) \geqslant N - 2M$ and an initial condition $x(0) \in \mathbb{R}^{n_x}$, such that

$$C_i e^{At} x(0) = 0, \, \forall i \in \mathcal{J}, \, t \in [0, T] \text{ and } x(0) \neq 0. \tag{7}$$

First note that if $N \leqslant 2M$, the empty set $\mathcal{J} \coloneqq \varnothing$ and an arbitrary non-zero initial condition satisfy (7). Henceforth, when (ii)' is false it is always true that there exists a set $\mathcal{J} \subset \{1, \ldots, N\}$ with $\text{card}(\mathcal{J}) \geqslant N - 2M$ and an initial condition $x(0) \in \mathbb{R}^{n_x}$, such that (7) holds. We shall prove that this contradicts (i). To this effect, select two disjoint sets $\mathcal{I}_a, \mathcal{I}_b \subset \{1, \ldots, N\}$, each with at most $M$ elements, so that $\mathcal{J} = \{1, 2, \ldots, N\} \backslash (\mathcal{I}_a \cup \mathcal{I}_b)$. Next define attack vectors $\eta = (\eta_1, \ldots, \eta_N) \in \mathcal{N}_{\mathcal{I}_a}$, $\bar{\eta} = (\bar{\eta}_1, \ldots, \bar{\eta}_N) \in \mathcal{N}_{\mathcal{I}_b}$ so that

$$\eta_i(t) \coloneqq -C_i e^{At} x(0), \forall i \in \mathcal{I}_a, \quad \bar{\eta}_i(t) \coloneqq C_i e^{At} x(0), \forall i \in \mathcal{I}_b,$$

$\forall t$, where $x(0)$ is the non-zero initial condition from (7). Since $\eta_i(t) = 0, \forall t \geqslant 0, i \notin \mathcal{I}_a$ and $\bar{\eta}_i(t) = 0, \forall t \geqslant 0, i \notin \mathcal{I}_b$, this choice for the attack vectors leads to

$$\begin{aligned} &C_i e^{At} x(0) + \eta_i(t) = 0, \; \bar{\eta}_i(t) = 0, && \forall i \in \mathcal{I}_a \\ &C_i e^{At} x(0) = \bar{\eta}_i(t), \; \eta_i(t) = 0, && \forall i \in \mathcal{I}_b \\ &C_i e^{At} x(0) = 0, \; \eta_i(t) = \bar{\eta}_i(t) = 0, \\ &\qquad \forall i \in \mathcal{J} = \{1, 2, \ldots, N\} \backslash (\mathcal{I}_a \cup \mathcal{I}_b), \end{aligned}$$

and therefore

$$C_i e^{At} x(0) + \eta_i(t) = \bar{\eta}_i(t), \; \forall i \in \{1, \ldots, N\}, t \in [0, T], \tag{8}$$

for some $x(0) \neq 0$. However, we can view the left-hand side expression $C_i e^{At} x(0) + \eta_i(t)$ as the output $y_i(t)$ associated with the initial condition $x(0) \neq 0$, the zero input, and the attack $\eta \in \mathcal{N}_{\mathcal{I}_a}$; whereas the right-hand side expression $\bar{\eta}_i(t)$ can be considered as the output $y_i(t)$ associated with the zero initial condition, zero input, and attack $\bar{\eta} \in \mathcal{N}_{\mathcal{I}_b}$. We have thus found two distinct initial conditions compatible with the same outputs, which contradicts observability under $M$ attacks and thus (i).

**(ii)' $\implies$ (i)**: Suppose by contradiction that (ii)' holds, but that (i) does not, and therefore that there exist initial conditions $x(0)$, $\bar{x}(0) \in \mathbb{R}^{n_x}$, an input $u(t)$, $t \geqslant 0$, sets $\mathcal{I}_a$, $\mathcal{I}_b \subset \{1, \ldots, N\}$ with at most $M$ elements, and attack

vectors $\eta = (\eta_1, \ldots, \eta_N) \in \mathcal{N}_{\mathcal{I}_a}$, $\bar{\eta} = (\bar{\eta}_1, \ldots, \bar{\eta}_N) \in \mathcal{N}_{\mathcal{I}_b}$ such that

$$y_i(t; x(0), u, \eta_i) = y_i(t; \bar{x}(0), u, \bar{\eta}_i), \quad \forall i \in \{1, \ldots, N\},\ t \in [0, T] \ \text{ and } \ x(0) \neq \bar{x}(0),$$

which, using the variation of constants formula, means that

$$C_i e^{At} x(0) + \mu_i(t) + \eta_i(t) = C_i e^{At} \bar{x}(0) + \mu_i(t) + \bar{\eta}_i(t), \quad \forall i \in \{1, \ldots, N\},\ t \in [0, T] \ \text{ and } \ x(0) \neq \bar{x}(0),$$

where $\mu_i(t) = C_i \int_0^t e^{A(t-s)} B u(s) ds$. Since $\eta_i(t) = 0$, $\forall t \geqslant 0, i \notin \mathcal{I}_a$ and $\bar{\eta}_i(t) = 0$, $\forall t \geqslant 0, i \notin \mathcal{I}_b$, we conclude that

$$C_i e^{At} \big(x(0) - \bar{x}(0)\big) = 0, \quad \forall i \in \mathcal{J},\ t \in [0, T] \ \text{ and } \ x(0) \neq \bar{x}(0),$$

where $\mathcal{J} \coloneqq \{1, 2, \ldots, N\} \backslash (\mathcal{I}_a \cup \mathcal{I}_b)$ is a set with no less than $N - 2M$ elements, which is in contradiction with (ii)'. Therefore, we have shown that (ii)' implies (i) and we have completed the proof. □

## III. Estimation algorithms

From Theorem 1, we know that an $N$-output system (1) is $M$-attack observable on $[0, T]$ if and only if for every subset $\mathcal{J}$ of $\{1, 2, \ldots, N\}$ with at least $N - 2M$ elements, the pair $(A, C_{\mathcal{J}})$ is observable. In this case, we can construct state estimators based on measurements from $N - 2M$ or more outputs on the interval $[0, T]$, which would provide accurate state estimates *in the presence of the attack signals* $\eta_i$.

An essential observation behind the design of the state estimators proposed here is that, for each combination of the $N - M$ (greater than $N - 2M$) outputs, we can construct one state estimator that would produce a correct state estimate based on measurements from those outputs in the interval $[0, T]$, in the absence of attacks $\eta_i$ on the chosen outputs. Moreover, assuming that at most $M$ sensors have been attacked, for each of these sets of $N - M$ outputs, there is at least one subset of $N - 2M$ outputs that consists of attack-free outputs. Hence, a state estimator based on this subset of $N - 2M$ outputs will result in an accurate estimate. We exploit this fact by proposing algorithms that choose wisely among several potential estimates to obtain good state estimates for the system (1). We propose an estimator that uses observability Gramians for state reconstruction in finite-time in Section III-A and an observer-based estimator in Section III-B, which we prove to be robust with respect to noise and disturbances.

### A. A Gramian-based estimator

Assume that the system (1) is $M$-attack observable on $[0, T]$. Given a set $\mathcal{J} \subset \{1, 2, \ldots, N\}$ with $N - 2M$ or more elements, we denote by $\hat{x}_{\mathcal{J}}(0)$ the initial state estimate produced by the observability Gramians using the input $u(t)$ and the outputs $y_i(t)$, for all $i \in \mathcal{J}$ collected in the interval $[0, T]$, that would be accurate if $\eta_i(t) = 0$, for every $i \in \mathcal{J}$ and $t \in [0, T]$. One can show that such estimate is given by

$$\hat{x}_{\mathcal{J}}(0) \ = \ W_{\mathcal{J}}(0, T)^{-1} \int_0^T e^{A^T s} C_{\mathcal{J}}^T \tilde{y}_{\mathcal{J}}(s) ds, \qquad (9)$$

where $\tilde{y}_{\mathcal{J}}(s) = y_{\mathcal{J}}(s) - \int_0^s C_{\mathcal{J}} e^{A(s-r)} B u(r) dr - D_{\mathcal{J}} u(s)$ (where $y_{\mathcal{J}}$ and $D_{\mathcal{J}}$ denotes the stacking of all $y_i$ and $D_i$ for $i \in \mathcal{J}$, respectively) and $W_{\mathcal{J}}(0, T) = \int_0^T e^{A^T s} C_{\mathcal{J}}^T C_{\mathcal{J}} e^{As} ds$ is the observability Gramian (see [4, Section 15.5]), which is invertible because the pair $(A, C_{\mathcal{J}})$ is observable (by Theorem 1).

For each subset $\mathcal{J} \subset \{1, 2, \ldots, N\}$ with $N - M$ ($\geqslant N - 2M$) elements, define $\pi_{\mathcal{J}}$ to be the largest deviation between the estimate $\hat{x}_{\mathcal{J}}(0)$ and any estimate that uses an $N - 2M$ subset $\mathcal{P} \subset \mathcal{J}$ of the outputs used to construct $\hat{x}_{\mathcal{J}}(0)$:

$$\pi_{\mathcal{J}} = \max_{\mathcal{P} \subset \mathcal{J} : \text{card}(\mathcal{P}) = N - 2M} |\hat{x}_{\mathcal{J}}(0) - \hat{x}_{\mathcal{P}}(0)|. \qquad (10)$$

When all the $\eta_i$, $i \in \mathcal{J}$ are equal to zero, all the estimates that appear in the definition of $\pi_{\mathcal{J}}$ will be consistent and we have $\pi_{\mathcal{J}} = 0$. This motivates the following state estimate:

$$\hat{x}(0) = \hat{x}_{\sigma}(0), \quad \sigma = \underset{\mathcal{J} \subset \{1, 2, \ldots, N\} : \text{card}(\mathcal{J}) = N - M}{\arg\min} \pi_{\mathcal{J}}. \qquad (11)$$

When more than one $\pi_{\mathcal{J}}$ achieve the minimum simultaneously, we can choose $\sigma$ to be any of them. We call this scheme a *finite-time Gramian-based estimator.* The following can be proved about this state estimator.

*Theorem 2:* Assume that the $N$-output system (1) is $M$-attack observable and that the attack vector $\eta$ belongs to $\mathcal{N}_{\mathcal{I}}$ for some set $\mathcal{I} \subset \{1, \ldots, N\}$ with $\text{card}(\mathcal{I}) \leqslant M$. For every initial conditions $x(0) \in \mathbb{R}^{n_x}$ and input $u$, the following holds

$$\hat{x}(0) = x(0), \qquad (12)$$

where $\hat{x}(0)$ is the estimate produced by the Gramian-based estimator (9)-(11). □

*Proof of Theorem 2.* Since system (1) is $M$-attack observable, we have from Theorem 1 that for any set $\mathcal{X} \subset \{1, 2, \ldots, N\}$ with $\text{card}(\mathcal{X}) \geqslant N - 2M$, the pair $(A, C_{\mathcal{X}})$ is observable. Following standard developments for Gramian-based reconstruction (see Section 15.6 of [4]), we rewrite the estimate $\hat{x}_{\mathcal{X}}(0)$ of the initial condition (9) in terms of the true initial condition $x(0)$ as follows

$$\begin{aligned}
\hat{x}_{\mathcal{X}}(0) &= W_{\mathcal{X}}(0, T)^{-1} \int_0^T e^{A^T s} C_{\mathcal{X}}^T \tilde{y}_{\mathcal{X}}(s) ds \\
&= W_{\mathcal{X}}(0, T)^{-1} \int_0^T e^{A^T s} C_{\mathcal{X}}^T C_{\mathcal{X}} x(s) ds - W_{\mathcal{X}}(0, T)^{-1} \\
&\quad \times \int_0^T e^{A^T s} C_{\mathcal{X}}^T \int_0^s C_{\mathcal{X}} e^{A(s-r)} B u(r) dr ds \\
&\quad + W_{\mathcal{X}}(0, T)^{-1} \int_0^T e^{A^T s} C_{\mathcal{X}}^T \eta_{\mathcal{X}}(s) ds \\
&= x(0) + W_{\mathcal{X}}(0, T)^{-1} \int_0^T e^{A^T s} C_{\mathcal{X}}^T \eta_{\mathcal{X}}(s) ds,
\end{aligned} \qquad (13)$$

where $\eta_{\mathcal{X}}$ denotes the stacking of all $\eta_i$, for $i \in \mathcal{X}$. We obtain the last equality since the first three (attack-free) terms reconstruct the true initial condition $x(0)$ exactly according to [4, Section 15.6]. Since $\eta = (\eta_1, \eta_2, \ldots, \eta_N) \in \mathcal{N}_{\mathcal{I}}$, we conclude from (13) with $\mathcal{X} = \bar{\mathcal{I}} \subset \{1, \ldots, N\} \backslash \mathcal{I}$

with $\text{card}(\bar{\mathcal{I}}) = N - M$ and also with $\mathcal{X} = \mathcal{P} \subset \bar{\mathcal{I}}$, $\text{card}(\mathcal{P}) = N - 2M$ that

$$\hat{x}_{\bar{\mathcal{I}}}(0) = \hat{x}_{\mathcal{P}}(0) = x(0) \tag{14}$$

which means that $\pi_{\bar{\mathcal{I}}} = 0$. Since $\pi_{\bar{\mathcal{I}}} = 0$ and $\sigma = \arg\min_{\mathcal{X}} \pi_{\mathcal{X}}$, we have that $\pi_\sigma = 0$ and therefore,

$$\hat{x}_\sigma(0) = \hat{x}_{\mathcal{P}}(0), \qquad \forall \mathcal{P} \subset \sigma : \text{card}(\mathcal{P}) = N - 2M. \tag{15}$$

Most importantly, since we are removing an additional $M$ elements from $\sigma$ to obtain the sets $\mathcal{P}$, regardless of what $\sigma$ turns out to be, there is always one set $\mathcal{P} \subset \sigma$, with $\text{card}(\mathcal{P}) = N - 2M$ for which $\eta_i(t) = 0$, for all $i \in \mathcal{P}$, $t \geqslant 0$. For this set $\hat{x}_{\mathcal{P}}(0) = x(0)$ and therefore we must necessarily have $\hat{x}_\sigma(0) = x(0)$, because of (15). □

Once we obtain an estimate of the initial condition $\hat{x}(0)$, we can then generate the state estimate for system (1) at any time $t \geqslant 0$ using

$$\hat{x}(t) = e^{At}\hat{x}(0) + \int_0^t e^{A(t-s)}Bu(s)ds. \tag{16}$$

Since we obtain $\hat{x}(0) = x(0)$ using the data $u(t)$ and $y(t)$ on the interval $[0, T]$, we achieve a correct estimate in *finite-time*, which is an advantage over the observer-based estimator introduced in the next section. However, the implementation of the Gramian-based estimator requires the inversion of the observability Gramians for each interval of time considered, which would be computationally very intensive if we wanted to construct a time series of state estimates. We will see that the observer-based estimator in the following section, only involves the solution of ordinary differential equations (ODEs), for which numerically efficient solvers are widely available.

### *B. An observer-based estimator*

We now consider an augmented version of system (1) with a process disturbance $d : [0, \infty] \to \mathbb{R}^{n_x}$ and measurement noise $m_i : [0, \infty] \to \mathbb{R}^{n_y}$, $i \in \{1, \ldots, N\}$, that enter the system in the following manner:

$$\begin{aligned} \dot{x} &= Ax + Bu + d \\ y_i &= C_i x + D_i u + \eta_i + m_i, \qquad i \in \{1, \ldots, N\}, \end{aligned} \tag{17}$$

Opposite to the attack signals $\eta_i$, all the measurement noise signals $m_i$ may be nonzero, but are typically bounded. Our goal is to show that the observer-based estimated proposed below is robust with respect to the process disturbance $d$ and the measurement noise $m_i$.

Following the same framework as the Gramian-based estimator in Section III-A, we assume that the $N$-output system (17) is observable through any $N - 2M$ outputs and construct an observer for every set $\mathcal{J} \subset \{1, \ldots, N\}$ with $N - M$ ($\geqslant N - 2M$) elements as follows:

$$\begin{aligned} \dot{\hat{x}}_{\mathcal{J}} &= A\hat{x}_{\mathcal{J}} + Bu + L_{\mathcal{J}}(\hat{y}_{\mathcal{J}} - y_{\mathcal{J}}) \\ \hat{y}_{\mathcal{J}} &= C_{\mathcal{J}}\hat{x}_{\mathcal{J}} + D_{\mathcal{J}}u, \end{aligned} \tag{18}$$

where the matrix $L_{\mathcal{J}}$ is chosen such that $A + L_{\mathcal{J}}C_{\mathcal{J}}$ is Hurwitz, which is always possible since every pair $(A, C_{\mathcal{J}})$ is observable (and therefore detectable) in view of Theorem 1.

From the bank of $\binom{N}{N-M}$ estimates $\hat{x}_{\mathcal{J}}$, we choose the state estimate along the lines followed by the Gramian-based estimator in Section III-A:

$$\hat{x}(t) = \hat{x}_{\sigma(t)}(t), \tag{19}$$

$$\sigma(t) = \underset{\mathcal{J} \subset \{1,2,\ldots,N\} : \text{card}(\mathcal{J}) = N-M}{\arg\min} \pi_{\mathcal{J}}(t), \tag{20}$$

$$\pi_{\mathcal{J}}(t) = \max_{\mathcal{P} \subset \mathcal{J} : \text{card}(\mathcal{P}) = N-2M} |\hat{x}_{\mathcal{J}}(t) - \hat{x}_{\mathcal{P}}(t)|, \tag{21}$$

where the state estimate $\hat{x}_{\mathcal{P}}$ for $\mathcal{P} \subset \mathcal{J}$ with $N - 2M$ elements is generated in the same manner as (18). The following result states that the proposed estimator is robust with respect to the disturbance $d$ and measurement noise $m_i$. For simplicity, we also initialize all the observers to the same condition $\hat{x}(0)$.

*Theorem 3:* Assume that the $N$-output system (1) is $M$-attack observable and $\eta_i$ in (17) belongs to $\mathcal{N}_{\mathcal{I}}$ for some set $\mathcal{I} \subset \{1, \ldots, N\}$ with $\text{card}(\mathcal{I}) \leqslant M$. There exist constants $\bar{k}$, $\bar{\alpha}$, $\bar{\gamma}_x$ and $\bar{\gamma}_y > 0$ such that for every initial condition $x(0) \in \mathbb{R}^{n_x}$ and input $u(t)$, $t \geqslant 0$, the following inequality holds along the trajectory of system (17):

$$\begin{aligned} |x(t) - \hat{x}(t)| \leqslant \bar{k} \exp(-\bar{\alpha}t)\, |x(0) - \hat{x}(0)| \\ + \bar{\gamma}_x \|d\|_{[0,t]} + \bar{\gamma}_y \left( \max_{\mathcal{J}} \|m_{\mathcal{J}}\|_{[0,t]} \right), \quad t \geqslant 0, \end{aligned} \tag{22}$$

for any initial conditions $x(0)$, $\hat{x}(0)$, $\hat{x}(0) \in \mathbb{R}^{n_x}$, as well as bounded signals $d$ and $m_i$, $i \in \{1, \ldots, N\}$, where we denote the stacking of all $m_i$, $i \in \mathcal{J}$ as $m_{\mathcal{J}}$. □

*Proof of Theorem 3* . For an arbitrary set $\mathcal{X} \subset \{1, \ldots, N\}$ with $\text{card}(\mathcal{X}) = N - 2M$ or $N - M$, the state estimation error $\tilde{x}_{\mathcal{X}} := x - \hat{x}_{\mathcal{X}}$ has the following error dynamics along solutions to the process (17) and the observer (18):

$$\dot{\tilde{x}}_{\mathcal{X}} = (A + L_{\mathcal{X}}C_{\mathcal{X}})\tilde{x}_{\mathcal{X}} - L_{\mathcal{X}}\eta_{\mathcal{X}} - L_{\mathcal{X}}m_{\mathcal{X}} + d. \tag{23}$$

Since $A + L_{\mathcal{X}}C_{\mathcal{X}}$ is Hurwitz and $\tilde{x}_{\mathcal{X}}(0) = x(0) - \hat{x}(0) = \tilde{x}(0)$ (as all observers are initialized at $\hat{x}(0)$ without loss of generality), the solution to (23) satisfies

$$\begin{aligned} |\tilde{x}_{\mathcal{X}}(t)| \leqslant\ & k_{\mathcal{X}} \exp(-\alpha_{\mathcal{X}}t)|\tilde{x}(0)| + \gamma_\eta \|\eta_{\mathcal{X}}\|_{[0,t]} \\ & + \gamma_y \|m_{\mathcal{X}}\|_{[0,t]} + \gamma_x \|d\|_{[0,t]}, \qquad \forall t \geqslant 0, \end{aligned} \tag{24}$$

where $k_{\mathcal{X}}$, $\alpha_{\mathcal{X}}$, $\gamma_\eta$, $\gamma_y$ and $\gamma_x > 0$. Since $\eta_i(t) = 0$, for all $i \in \{1, \ldots, N\} \backslash \mathcal{I}$ and $t \geqslant 0$, we conclude from (24) with $\mathcal{X} = \bar{\mathcal{I}} \subseteq \{1, \ldots, N\} \backslash \mathcal{I}$ with $\text{card}(\bar{\mathcal{I}}) = N - M$ that

$$\begin{aligned} |\tilde{x}_{\bar{\mathcal{I}}}(t)| \leqslant\ & k_{\bar{\mathcal{I}}} \exp(-\alpha_{\bar{\mathcal{I}}}t)|\tilde{x}(0)| + \gamma_y \|m_{\bar{\mathcal{I}}}\|_{[0,t]} \\ & + \gamma_x \|d\|_{[0,t]}, \qquad t \geqslant 0. \end{aligned} \tag{25}$$

and also for any set $\mathcal{P} \subset \bar{\mathcal{I}}$ with $\text{card}(\mathcal{P}) = N - 2M$, we have from (24) with $\mathcal{X} = \mathcal{P}$ that

$$\begin{aligned} |\tilde{x}_{\mathcal{P}}(t)| \leqslant\ & k_{\mathcal{P}} \exp(-\alpha_{\mathcal{P}}t)|\tilde{x}(0)| + \gamma_y \|m_{\mathcal{P}}\|_{[0,t]} \\ & + \gamma_x \|d\|_{[0,t]}, \qquad t \geqslant 0. \end{aligned} \tag{26}$$

Recalling the definition of $\pi_{\bar{\mathcal{I}}}$ from (10), we have that

$$\begin{aligned}\pi_{\bar{\mathcal{I}}}(t) &= \max_{\mathcal{P}\subset\bar{\mathcal{I}}} |\hat{x}_{\bar{\mathcal{I}}}(t) - \hat{x}_{\mathcal{P}}(t)| \\ &= \max_{\mathcal{P}\subset\bar{\mathcal{I}}} |\hat{x}_{\bar{\mathcal{I}}}(t) - x(t) + x(t) - \hat{x}_{\mathcal{P}}(t)| \\ &\leqslant |\tilde{x}_{\bar{\mathcal{I}}}(t)| + \max_{\mathcal{P}\subset\bar{\mathcal{I}}} |\tilde{x}_{\mathcal{P}}(t)|\,.\end{aligned} \tag{27}$$

From (25) and (26), we obtain

$$\begin{aligned}\pi_{\bar{\mathcal{I}}}(t) \leqslant 2k\exp(-\alpha t)|\tilde{x}_{\bar{\mathcal{I}}}(0)| + 2\gamma_y\|m_{\bar{\mathcal{I}}}\|_{[0,t]} \\ + 2\gamma_x\|d\|_{[0,t]}, \qquad t \geqslant 0,\end{aligned} \tag{28}$$

where $k := \max_{\mathcal{P}\subset\bar{\mathcal{I}}} \{k_{\bar{\mathcal{I}}}, k_{\mathcal{P}}\}$ and $\alpha := \min_{\mathcal{P}\subset\bar{\mathcal{I}}} \{\alpha_{\bar{\mathcal{I}}}, \alpha_{\mathcal{P}}\}$. Observe that for every $\mathcal{J}$ with $\text{card}(\mathcal{J}) = N - M$, we have at least one set $\bar{\mathcal{P}} \subset \mathcal{J}$ with $\text{card}(\bar{\mathcal{P}}) = N - 2M$ satisfying

$$\begin{aligned}|\tilde{x}_{\bar{\mathcal{P}}}(t)| \leqslant k\exp(-\alpha t)|\tilde{x}_{\bar{\mathcal{P}}}(0)| \\ +\gamma_y\|m_{\bar{\mathcal{P}}}\|_{[0,t]} + \gamma_x\|d\|_{[0,t]},\ t \geqslant 0.\end{aligned} \tag{29}$$

Recall from (21) that $\hat{x}(t) = \hat{x}_{\sigma(t)}(t)$ where $\sigma(t) = \underset{\mathcal{J}:\text{card}(\mathcal{J})=N-M}{\arg\min} \pi_{\mathcal{J}}(t)$, hence $\pi_{\sigma(t)}(t) \leqslant \pi_{\bar{\mathcal{I}}}(t)$. Using the fact that $\pi_{\sigma(t)}(t) := \max_{\mathcal{P}\subset\sigma:\text{card}(\mathcal{P})=N-2M} |\hat{x}_{\sigma(t)}(t) - \hat{x}_{\mathcal{P}}(t)| \geqslant |\hat{x}_{\sigma(t)}(t) - \hat{x}_{\bar{\mathcal{P}}}|$, we have from the triangle inequality that

$$\begin{aligned}|x(t) - \hat{x}_{\sigma(t)}(t)| &= |\tilde{x}_{\sigma(t)}(t)| \\ &= |x(t) - \hat{x}_{\bar{\mathcal{P}}}(t) + \hat{x}_{\bar{\mathcal{P}}}(t) - \hat{x}_{\sigma(t)}(t)| \\ &\leqslant |\tilde{x}_{\bar{\mathcal{P}}}(t)| + |\hat{x}_{\bar{\mathcal{P}}}(t) - \hat{x}_{\sigma(t)}(t)| \\ &\leqslant |\tilde{x}_{\bar{\mathcal{P}}}(t)| + \pi_{\sigma(t)}(t) \\ &\leqslant |\tilde{x}_{\bar{\mathcal{P}}}(t)| + \pi_{\bar{\mathcal{I}}}(t), \qquad t \geqslant 0.\end{aligned} \tag{30}$$

From (28) and (29), we have

$$\begin{aligned}|\tilde{x}_{\sigma(t)}(t)| \leqslant 3k\exp(-\alpha t)|\tilde{x}(0)| \\ +3\gamma_y\left(\max\{\|m_{\bar{\mathcal{P}}}\|_{[0,t]}, \|m_{\bar{\mathcal{I}}}\|_{[0,t]}\}\right) \\ +3\gamma_x\|d\|_{[0,t]}, \qquad t \geqslant 0.\end{aligned} \tag{31}$$

We see that (31) satisfies (22) by setting $\bar{k} := 3k$, $\bar{\alpha} := \alpha$, $\bar{\gamma}_y := 3\gamma_y$ and $\bar{\gamma}_x := 3\gamma_x$, which concludes the proof. □

The proposed observer-based estimator provides exponential convergence of the estimates to a neighborhood of the true states $x$ under the assumption that the perturbed system (17) is $M$-attack observable. In other words, the robust observer (18)-(21) generates an error system that is input-to-state stable (ISS) according to the definition of [12] with respect to the process disturbance $d$ and output measurement noises $m_i$. When there are no disturbances, we obtain exponential convergence of the estimates to the true states for all initial conditions.

*Remark 3:* The observer-based estimator only requires detectability as opposed to observability in the Gramian-based estimator, which is counterpointed by asymptotic, instead of finite-time convergence of the states. □

## IV. Conclusions

We introduced a new notion of observability for multi-output continuous-time LTI systems, for which a subset of the outputs can be attacked by an adversary. A necessary and sufficient condition is derived which allows standard observability tests to be employed in checking whether a system is 'observable under attacks'. We propose two state-estimation algorithms: a finite-time Gramian-based estimator and an asymptotic observer-based estimator. For the latter, we show that it provides bounded estimation errors in an ISS-like manner in the presence of bounded disturbances and measurement noise. Future works include the consideration of the stabilization problem and reducing the computational complexity of the proposed estimation algorithms.

## References


[1] S. Amin, A.A. Cárdenas, and S.S. Sastry. Safe and secure networked control systems under denial-of-service attacks. In *Hybrid Systems: Computation and Control*, pages 31–45. Springer, 2009.

[2] H. Fawzi, P. Tabuada, and S. Diggavi. Secure estimation and control for cyber-physical systems under adversarial attacks. *IEEE Transactions on Automatic Control*, 59(6):1454–1467, June 2014.

[3] J.M. Hendrickx, K.H. Johansson, R.M. Jungers, H. Sandberg, and K.C. Sou. Efficient computations of a security index for false data attacks in power networks. *IEEE Transactions on Automatic Control*, 59(12):3194–3208, December 2014.

[4] J.P. Hespanha. *Linear systems theory*. Princeton University Press, 2009.

[5] M. Jones, G. Kotsalis, and J.S. Shamma. Cyber-attack forecast modeling and complexity reduction using a game-theoretic framework. In *Control of Cyber-Physical Systems*, pages 65–84. Springer, 2013.

[6] Y. Liu, P. Ning, and M.K. Reiter. False data injection attacks against state estimation in electric power grids. *ACM Transactions on Information and System Security (TISSEC)*, 14(1):13, 2011.

[7] M.H. Manshaei, Q. Zhu, T. Alpcan, T. Basar, and J. Hubaux. Game theory meets network security and privacy. *ACM Computing Surveys (CSUR)*, 45(3):25, 2013.

[8] M.A. Massoumnia, G.C. Verghese, and A.S. Willsky. Failure detection and identification. *IEEE Transactions on Automatic Control*, 34(3):316–321, March 1989.

[9] Y. Mo, J.P. Hespanha, and B. Sinopoli. Resilient detection in the presence of integrity attacks. *IEEE Transactions on Signal Processing*, 62(1):31–43, January 2014.

[10] F. Pasqualetti, F. Dorfler, and F. Bullo. Attack detection and identification in cyber-physical systems. *IEEE Transactions on Automatic Control*, 58(11):2715–2729, November 2013.

[11] Y. Shoukry and P. Tabuada. Event-triggered projected luenberger observer for linear systems under sparse sensor attacks. In *Proccedings of the 53rd IEEE Conference on Decision and Control (CDC)*, 2014.

[12] E.D. Sontag. Input to state stability: Basic concepts and results. *Nonlinear and Optimal Control Theory*, 1932:163–220, 2008.

[13] A. Teixeira, S. Amin, H. Sandberg, K.H. Johansson, and S.S. Sastry. Cyber security analysis of state estimators in electric power systems. In *Proccedings of the 49th IEEE Conference on Decision and Control (CDC)*, pages 5991–5998, 2010.

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