IEEE Control Systems Letters paper presented at
2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Secure State Estimation under Actuator and Sensor Attacks using Sliding Mode Observers

Twan Keijzer[1], Riccardo M.G. Ferrari[1], Henrik Sandberg[2]

***Abstract*— Interconnections in modern systems make them vulnerable to adversarial attackers both by corrupting communication channels and compromising entire subsystems. The field of secure state estimation (SSE) aims to provide correct state estimation even when an unknown part of the measurement signals is corrupted. In this paper, we propose a solution to a novel generalized SSE problem in which full subsystems can be compromised, corrupting both the actuation and measurement signals. For a full system with p measurements, the proposed sliding mode observer (SMO)-based solution allows for up to p attack channels which can be arbitrarily distributed amongst attacks on actuation and measurement signals. This is a much larger class of attacks than considered in the existing literature. The method is demonstrated on 10 interconnected mass-spring-damper subsystems.**

I. Introduction

The abundance of interconnections in modern control systems has allowed them to provide better performance and be more resilient against faults. It however also exposes these systems to adversarial attackers who can corrupt data on communication channels or even compromise entire subsystems [1]. While it is often impossible to know which subsystems will be targeted, an attacker is typically assumed to have limited resources [2]. A logical way to quantify this attacker limitation is to assume an upper bound on the maximum number of subsystems they can compromise.

To this end, inspired by the Byzantine general's problem [3], the field of secure state estimation (SSE) aims to provide correct state estimation when an unknown, but limited, part of the measurements is corrupted [4], [5], [6], [7], [8], [9]. This secure state estimate allows for cyber-attack tolerant control while retaining the nominal controller. Such SSE methods have been developed to be applicable to nonlinear systems [6], or to be implemented in a distributed fashion [10] or with reduced computational complexity [9]. Furthermore, [8] proposes a set-based SSE method that allows for all but one measurement to be compromised. These methods are, however, all limited to attacks on the measurements.

Literature on generalized SSE problems which also address actuator attacks is limited. In [11], [12] an optimization-based decoder is proposed to estimate actuator attacks. The number of allowed attacks for this approach is however limited to strictly less than half the number of measurements.

[1]Delft Center for Systems and Control, Delft University of Technology, The Netherlands. {t.keijzer, r.Ferrari}@tudelft.nl.
[2]School of Electrical Engineering and Computer Science and Digital Futures, KTH Royal Institute of Technology, Stockholm, Sweden. hsan@kth.se.
Thanks to Alexander J. Gallo for the fruitful discussions leading to the proof of Lemma 4.

A sliding mode observer (SMO)-based approach is proposed in [13], allowing for attacks on a known subset of actuators when corresponding measurements are uncompromised.

In this paper, we consider a system of subsystems that can be interconnected physically or via a distributed control law. For such a system we solve a generalized SSE problem where attacked subsystems are fully compromised, i.e. both the actuator and measurements of affected subsystems are corrupted. Similar to the traditional SSE we allow for an unknown, but limited, part of the subsystems to be fully compromised. SMOs are well suited to address this generalized SSE problem as they can estimate the state as well as anomalies acting on the actuators or sensors [14], [15].

The main contribution of this paper, therefore, is an SMO-based solution to the generalised SSE problem which can tolerate as many simultaneous attacks as the number of measurements. This represents a much larger class of attacks than is currently addressed in the literature [8], [11], [12]. To achieve this, we design a bank of SMOs that can be used to both isolate which subsystems are compromised, and to estimate the attacks affecting them. In particular, an SMO is designed for each possible hypothesis of which subsystem is attacked and which is not. We then leverage the so-called matching condition to cross-validate the SMOs estimates against each other and isolate the correct hypotheses.

In the remainder of this paper we introduce the problem in section II. In section III the bank of SMO-based state and attack estimators are designed. The main contribution of this work, namely a method to use the SMO-based attack estimates to isolate the compromised subsystems, is presented in section IV. In section V the method is demonstrated in simulation. The work is concluded in section VI.

A. Notation

For a set $\mathcal{N}$ let us denote by $|\mathcal{N}|$ the cardinality of the set. $\binom{n}{k}$ denotes the binomial coefficient 'n choose k'. $\lceil x \rceil$ denotes rounding up to the next integer. $\mathbb{C}^-$ denotes the set of complex numbers with negative real part. Lastly, denote the time series from 0 to t of a variable x by $x[0:t]$ and denote $\mathbf{0}$ as a time series of only zeros.

II. Problem Statement

Consider a set of N linear interconnected subsystems as

$$\begin{cases} \dot{x}_\ell = A_\ell x_\ell + B_\ell f_{u_\ell} + \sum_{j \in \mathcal{N}} A_{\ell j} x_j \,, \\ y_\ell = C_\ell x_\ell + D_\ell f_{y_\ell} + \sum_{j \in \mathcal{N}} C_{\ell j} x_j \,, \end{cases} \tag{1}$$

where $x_\ell \in \mathbb{R}^{n_\ell}$, $y_\ell \in \mathbb{R}^{p_\ell}$, $f_{u_\ell} \in \mathbb{R}^{m_\ell}$, $f_{y_\ell} \in \mathbb{R}^{p_\ell}$ and $\mathcal{N} = \{1, \ldots, N\}$ is the set of all subsystems. The matrices $A_{\ell j}$ account either for physical interconnections or the presence of a distributed state-feedback control law. This gives a global system dynamics as

$$\begin{cases} \dot{x} = Ax + Bf_u \,, \\ y = Cx + Df_y \,, \end{cases} \tag{2}$$

where B and D are block-diagonal and full column rank, $x \in \mathbb{R}^n$, $y \in \mathbb{R}^p$, $f_u \in \mathbb{R}^m$, $f_y \in \mathbb{R}^p$. Note here that $m = \sum_{\ell \in \mathcal{N}} m_\ell$ represents the number of possible input attacks and $p = \sum_{\ell \in \mathcal{N}} p_\ell$ represents both the number of outputs and possible output attacks.

Remark 1: A known input u can be added without affecting the results. It has been omitted to simplify notation. ◁

We consider that this system is subject to attacks that can compromise an unknown subset of subsystems $\mathcal{A} \subset \mathcal{N}$. If a subsystem is compromised, we consider both f_{u_ℓ} and f_{y_ℓ} to be potentially non-zero. Furthermore, both the attacker and defender are considered to have full model knowledge and full disclosure resources. We make the following assumptions about the system and attack.

Assumption 1: $\mathcal{A}$ is unknown, but constant over time. ◁

Note that the set $\mathcal{A}$ not being known differentiates the generalised secure state estimation problem solved in this paper from state estimation with unknown inputs as researched in [14], [16], [17] and many others.

Assumption 2: The system in (2) is state and input observable. [18] ◁

Proposition 1: The total number of active attacks is at most equal to the number of measurements, i.e.

$$\sum_{\ell \in \mathcal{A}} (m_\ell + p_\ell) \leq p \,,$$

is a necessary condition for Assumption 2 to hold.

Proof: The system is state and input observable by Assumption 2, therefore the proof follows directly from Corollary 1 in [18]. ■

Assumption 3: If $m \neq 0$, the full system state x is observable from any combination of $N - |\mathcal{A}|$ outputs y_ℓ. ◁

Remark 2: Assumption 3 requires the subsystems to be sufficiently interconnected. This is necessary to estimate the attacks on the input from the un-attacked outputs. Note that if $m = 0$ there are no attacks on the input and thus Assumption 3 is not required. Assumption 2 provides state and input observability. Note that, except for the limit case where $m = 0$, Assumption 3 implies the state observability claim. The input observability claim of 2 is required to prevent the existence of zero dynamics attacks [19] which can lead to wrong attack identification. If zero dynamics attacks do not need to be prevented assumption 2 can be relaxed to $m_\ell < n_\ell$, $\forall \ell$. Lastly, Proposition 1 and Assumption 1 define the limitations on the attacker resources. ◁

The considered problem is a generalization of the SSE problem considering attacks on both input and output. Furthermore, the presented approach allows for p signals to be attacked. To the author's best knowledge, current approaches allow for at most $p - 1$ measurement signals [8] or $\lceil \frac{p-1}{2} \rceil$ measurement and input signals [11], [12] to be attacked.

Remark 3: Assumption 2 is common in literature on SSE where also inputs are subject to attack [11], [12]. Furthermore, assumption 3, which implies state observability, is common in literature on the standard SSE problem, see for example Assumption 1.ii in [8] or implicitly in [4]. Assumption 1 also appears in several works such as [11], [12], but is not required in [8]. Proposition 1 is a relaxation of the common assumption that the number of attacks is strictly less than half of the number of outputs. ◁

III. Sliding Mode Observer Design

In this section we will design a bank of SMOs where each is suited to a different attack scenario. We will prove that for every possible attack scenario, there exists an SMO that provides a correct state and attack estimate. We will then present a method to identify the correct SMOs in Section IV.

To this end, let us introduce the set $\mathcal{I} = \{1, \ldots, I\}$, enumerating all the possible hypotheses on the composition of the set $\mathcal{A}$ of attacked subsystems. Based on each hypothesis $i \in \mathcal{I}$ we design a sliding mode observer (SMO) that provides a correct state and attack estimate under that hypothesis. This is possible due to the capability of SMOs to reject matched anomalies [14]. We will denote the set of correct hypotheses as $I_\mathcal{A} \subseteq \mathcal{I}$, with $I_\mathcal{A} \neq \emptyset$ by definition. Below we present an example to clarify the set definitions.

Example 1: Consider $\mathcal{N} = \{1, 2, 3\}$ and $p_\ell = 2$, $m_\ell = 1 \; \forall \ell$. Then, by proposition 1 at most 2 subsystems can be attacked, i.e. $\mathcal{A}$ can be $\{1, 2\}$, $\{1, 3\}$, $\{2, 3\}$, $\{1\}$, $\{2\}$, $\{3\}$, $\emptyset$. Based on the possible attacks we choose hypotheses $\mathcal{A} = \{1, 2\}$, $\mathcal{A} = \{1, 3\}$, $\mathcal{A} = \{2, 3\}$ to design the SMOs. These hypotheses are enumerated in $\mathcal{I} = \{1, 2, 3\}$. If in fact $\mathcal{A} = 2$ then hypotheses 1 and 3 provide correct state and attack estimates and thus $\mathcal{I}_\mathcal{A} = \{1, 3\}$.

The number of possible hypotheses I can be upper bounded as $I \leq \binom{N}{\max_{|\mathcal{A}|} |\mathcal{A}|}$, which reduces to an equality if all subsystems have the same number of inputs and the same number of outputs. To obtain these SMOs we first transform the system in equation (2) in a way that allows to distinguish the attacked and healthy subsystems. For a given hypothesis i this transformation is defined as $\begin{bmatrix} \tilde{x}_1^i \\ \tilde{x}_2^i \end{bmatrix} = T_x^i x$, $\begin{bmatrix} \tilde{y}_2^i \\ \tilde{y}_2^i \end{bmatrix} = T_y^i y$, $\begin{bmatrix} f_{y_1}^i \\ f_{y_2}^i \end{bmatrix} = T_{f_y}^i f_y$, and $\begin{bmatrix} f_{u_1}^i \\ f_{u_2}^i \end{bmatrix} = T_u^i f_u$ which gives the transformed system

$$\begin{cases} \begin{bmatrix} \dot{\tilde{x}}_1^i \\ \dot{\tilde{x}}_2^i \end{bmatrix} = \underbrace{\begin{bmatrix} A_{11}^i & A_{12}^i \\ A_{21}^i & A_{22}^i \end{bmatrix}}_{A^i} \begin{bmatrix} \tilde{x}_1^i \\ \tilde{x}_2^i \end{bmatrix} + \underbrace{\begin{bmatrix} B_{11}^i & 0 \\ 0 & B_{22}^i \end{bmatrix}}_{B^i} \begin{bmatrix} f_{u_1}^i \\ f_{u_2}^i \end{bmatrix} \,, \\ \begin{bmatrix} \tilde{y}_1^i \\ \tilde{y}_2^i \end{bmatrix} = \underbrace{\begin{bmatrix} C_{11}^i & C_{12}^i \\ C_{21}^i & C_{22}^i \end{bmatrix}}_{C^i} \begin{bmatrix} \tilde{x}_1^i \\ \tilde{x}_2^i \end{bmatrix} + \begin{bmatrix} D_{11}^i & 0 \\ 0 & D_{22}^i \end{bmatrix} \begin{bmatrix} f_{y_1}^i \\ f_{y_2}^i \end{bmatrix} \,. \end{cases} \tag{3}$$

The transformations just introduced are such that $f_{u_1}^i = f_{y_1}^i = 0 \; \forall t$ if hypothesis i is correct, i.e. if $i \in \mathcal{I}_\mathcal{A}$. Note that this transformation is only possible due to the block-diagonal structure of B and D, which appears as a result of the subsystems definition in (1). Therefore, in the remainder of this section, we will design an SMO which produces a correct state and attack estimate if $f^i_{u_1} = f^i_{y_1} = 0\ ,\forall t$. To be able to design such an SMO the system first needs to be manipulated to adhere to the following two conditions that are common for SMOs [16]:

1) Non-minimum phase condition: The invariant zeros of (A^i, B^i, C^i) lie in $\mathbb{C}^-$.
2) Matching condition: relative degree between attack and output is 1.

The non-minimum phase condition is implied by Assumption 2. However, the matching condition does not trivially hold for all $i \in \mathcal{I}$. In sections III-A and III-B we present two extensions of the system with which we can guarantee the matching condition to hold for all $i \in \mathcal{I}$.

While *naively* designing the SMO to provide correct state and attack estimates in the case $i \in \mathcal{I}_\mathcal{A}$, we will also keep track of the effect of $f^i_{u_1}$ and $f^i_{y_1}$ on the SMO state and attack estimates when $i \notin \mathcal{I}_\mathcal{A}$. In section IV, we will then show how the combination of all I observers can be used to identify the set $\mathcal{I}_\mathcal{A}$ and for state and attack estimation.

A. Extend the System with Filtered Measurements

Part of the attack that is to be estimated, f_{y_2}, directly affects the output, i.e. has relative degree 0 with respect to the output. One can make the matching condition hold in such cases by filtering the affected outputs as in [14]. By applying this approach to the system in (3) we obtain

$$\begin{cases} \begin{bmatrix}\dot{\tilde{x}}_1\\ \dot{\tilde{x}}_2\\ \dot{z}\end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & 0\\ A_{21} & A_{22} & 0\\ -A_fC_{21} & -A_fC_{22} & A_f\end{bmatrix}\begin{bmatrix}\tilde{x}_1\\ \tilde{x}_2\\ z\end{bmatrix} \\ \qquad + \begin{bmatrix} B_{11} & 0 & 0\\ 0 & B_{22} & 0\\ 0 & 0 & -A_fD_{22}\end{bmatrix}\begin{bmatrix} f_{u_1}\\ f_{u_2}\\ f_{y_2}\end{bmatrix}, \\ \begin{bmatrix}\tilde{y}_1\\ z\end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} & 0\\ 0 & 0 & I\end{bmatrix}\begin{bmatrix}\tilde{x}_1\\ \tilde{x}_2\\ z\end{bmatrix} + \begin{bmatrix} D_{11}\\ 0\end{bmatrix} f^i_{y_1}, \end{cases} \tag{4}$$

where we omitted superscript i to ease notation. Here A_f is a full-rank Hurwitz matrix and z is the filtered $\tilde{y}_2$. One can see that the matching condition now holds for f_{y_2} by design.

B. Extend the Output using HOSM Differentiators

For the matching condition to hold for f_{u_2} we require it to be relative degree 1 with respect to the output. This might not inherently be the case for all $i \in \mathcal{I}$. Therefore, if necessary, we extend the output with derivatives obtained from the higher order sliding mode (HOSM) differentiator from [20]. This gives the extended output

$$\underbrace{\begin{bmatrix}\tilde{y}_1\\ \tilde{y}_d\\ z\end{bmatrix}}_{\tilde{y}_e} = \begin{bmatrix} C_{11} & C_{12} & 0\\ \tilde{C}_{11} & \tilde{C}_{12} & 0\\ 0 & 0 & I\end{bmatrix}\begin{bmatrix}\tilde{x}_1\\ \tilde{x}_2\\ z\end{bmatrix} + \begin{bmatrix} D_{11} & 0\\ 0 & I\\ 0 & 0\end{bmatrix}\begin{bmatrix} f^i_{y_1}\\ \epsilon\end{bmatrix}, \tag{5}$$

where $[\tilde{C}_{11}\tilde{C}_{12}]$ are rows of the matrix obtained from the product $[C_{11}C_{12}]\left[(A^i)^\top \quad (A^{i^2})^\top \quad \dots \quad (A^{i^n})^\top\right]^\top$ such that $\text{rank}(\begin{bmatrix}C_{12}\\ \tilde{C}_{12}\end{bmatrix} B_2) = \text{rank}(B_2)$, i.e. the relative degree is 1 and the matching condition holds. This can always be achieved due to Assumption 3. Furthermore, ϵ is the error of the HOSM differentiator, which is a function of $f^i_{y_1}[0:t]$ and is defined in Equation (6) of [20]. For the remainder of this paper, the only relevant properties of $\epsilon(f^i_{y_1}[0:t])$ is that $\epsilon(0) = 0$ and it is bounded for bounded input.

C. Transform system to the SMO standard form

Now that matching and non-minimum phase conditions hold, the extended system can be transformed into the SMO standard form [16], [21][1]. To this end we perform the transformations $y_e = T_{y_e}\tilde{y}_e$, $\begin{bmatrix}x_1\\ x_2\end{bmatrix} = T_{xz}\begin{bmatrix}\tilde{x}_1\\ \tilde{x}_2\\ z\end{bmatrix}$ and $\begin{bmatrix}f_1\\ f_{u_{12}}\end{bmatrix} = T_{f_{u_1}}f_{u_1}$, and define $f_2 = \begin{bmatrix}f_{u_{12}}\\ f_{u_2}\\ f_{y_2}\end{bmatrix}$ and $f_3 = \begin{bmatrix}f_{y_1}\\ \epsilon_1\end{bmatrix}$ to obtain the SMO standard form as

$$\begin{cases} \begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix} = \begin{bmatrix}A_1 & A_2\\ A_3 & A_4\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} + \begin{bmatrix}B_1 & 0\\ B_{21} & B_2\end{bmatrix}\begin{bmatrix}f_1\\ f_2\end{bmatrix}, \\ \quad y_e = \begin{bmatrix}0 & I\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} + Df_3\,. \end{cases} \tag{6}$$

Here $T_{f_{u_1}}$ is chosen such that the matching condition does not hold for attacks in f_1 and f_2 contains only attacks for which the matching condition does hold. Note that as a result of these transformations $f_1 = f_3 = 0\ ,\forall t$ and $\forall i \in \mathcal{I}_\mathcal{A}$.

D. SMO for state and attack estimation

For the system in (6) an SMO can be designed as

$$\begin{cases} \begin{bmatrix}\dot{\hat{x}}_1\\ \dot{\hat{x}}_2\end{bmatrix} = \begin{bmatrix}A_1 & A_2\\ A_3 & A_4\end{bmatrix}\begin{bmatrix}\hat{x}_1\\ \hat{x}_2\end{bmatrix} + \begin{bmatrix}-A_2\\ G_l\end{bmatrix} e_y + \begin{bmatrix}0\\ \nu\end{bmatrix}, \\ \hat{y}_e = \hat{x}_2\,, \\ \nu = -\rho\,\text{sign}(Pe_y)\,, \end{cases} \tag{7}$$

where $G_l = A_s - A_4$, A_s is Hurwitz, and $e_y = \hat{y}_e - y_e$. Thus the state estimation error dynamics can be written as

$$\begin{cases} \begin{bmatrix}\dot{e}_1\\ \dot{e}_2\end{bmatrix} = \begin{bmatrix}A_1 & 0\\ A_3 & A_s\end{bmatrix}\begin{bmatrix}e_1\\ e_2\end{bmatrix} - \begin{bmatrix}B_1 & 0 & -A_2D\\ B_{21} & B_2 & G_lD\end{bmatrix}\begin{bmatrix}f_1\\ f_2\\ f_3\end{bmatrix} + \begin{bmatrix}0\\ \nu\end{bmatrix}, \\ e_y = e_2 - Df_3\,, \end{cases} \tag{8}$$

with $e_1 = \hat{x}_1 - x_1$ and $e_2 = \hat{x}_2 - x_2$.

Lemma 1: e_1 and e_2 converge to 0 for any $i \in \mathcal{I}_\mathcal{A}$ if $\rho > \max_t ||B_2f_2(t)||$

Proof: Recall that $f_1 = f_3 = 0\ ,\forall t$ for any $i \in \mathcal{I}_\mathcal{A}$. The proof then follows from Proposition 2 in [14]. ■

Note that in most systems very large attacks are trivially detected. Therefore ρ can be chosen accordingly without knowledge of the actual attack.

Corollary 1: If there are only attacks on the measurements, the state can be securely estimated with a single SMO and without the need for attack isolation.

Proof: If there are only attacks on the measurements, $m_\ell = 0\ ,\forall \ell$, Assumption 2 holds trivially even for $\mathcal{A} = \mathcal{N}$. Therefore, $|\mathcal{I}| \le \binom{N}{|\mathcal{N}|} = 1$ and $\mathcal{I}_\mathcal{A} = \mathcal{I}$. ■

[1]The SMO standard form does not consider the additional attacks f_{u_1} and f_{y_1} that do no adhere to the matching condition. The standard form presented here is a generalization of the standard form that does.

For $i \notin \mathcal{I}_\mathcal{A}$ the state estimation behaviour is analyzed below.

Lemma 2: For $\rho > \max_t ||A_3 e_1 + A_4 D f_3 - B_{21} f_1 - B_2 f_2 - D\dot{f}_3||$, the sliding surface $e_2 = Df_3$ is reached in finite time.

Proof: It can be shown that the sliding motion on sliding surface $e_y = 0$ will take place in finite time using the same approach as the proof of Proposition 2 in [14]. By (8) the sliding surface $e_y = 0$ is equivalent to $e_2 = Df_3$. ■

Remark 4: If ρ is chosen only as $\rho > \max_t ||B_2 f_2||$ then for $i \notin \mathcal{I}_\mathcal{A}$ we might not reach the sliding surface $e_y = 0$. Therefore, if the ρ as in Lemma 2 becomes excessively large, $e_y = 0$ can be used as an additional condition to identify a correct hypothesis of the attack. ◁

Substituting $e_2 = Df_3$ and $\dot{e}_2 = D\dot{f}_3$ into (8), obtain ν_{eq} as

$$\begin{cases} \begin{bmatrix} \dot{e}_1 \\ \dot{f}_3 \end{bmatrix} = \begin{bmatrix} A_1 & A_2 D \\ 0 & 0 \end{bmatrix} \begin{bmatrix} e_1 \\ f_3 \end{bmatrix} + \begin{bmatrix} -B_1 & 0 & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ \dot{f}_3 \end{bmatrix}, \\ \nu_{\text{eq}} = [-A_3 \quad -A_4 D] \begin{bmatrix} e_1 \\ f_3 \end{bmatrix} + [B_{21} \quad B_2 \quad D] \begin{bmatrix} f_1 \\ f_2 \\ \dot{f}_3 \end{bmatrix}. \end{cases} \tag{9}$$

Typically, the attack estimate is then calculated as $B_2^\dagger \nu_{\text{eq}}$ [14]. In this work, however, we will calculate $\hat{f} = \begin{bmatrix} \hat{f}_2 \\ \hat{f}_r \end{bmatrix} = \begin{bmatrix} B_2^\dagger \\ T_B \end{bmatrix} \nu_{\text{eq}}$ where $T_B^\top$ spans the null space of $B_2^\dagger$.

Remark 5: The additional rows of the attack estimate are not useful for attack estimation if $i \in \mathcal{I}_\mathcal{A}$, but are added to make sure no information about the attack is lost if $i \notin \mathcal{I}_\mathcal{A}$. This is required for the attack identification algorithm presented in section IV. ◁

Lemma 3: $\hat{f}_2 = f_2$ for any $i \in \mathcal{I}_\mathcal{A}$ after the sliding surface is reached.

Proof: Given assumption 2, B_2 is full column rank and the proof follows directly from [14]. ■

In the next section, we will use the attack estimate $\hat{f}$ for attack identification. To simplify notation, we will denote the following function

$$\hat{f}^i = f^i_{\text{est}}(f^i_1[0:t], f^i_2, f^i_3[0:t]), \tag{10}$$

where $[0:t]$ denotes the full time-series of the attack. Note that from this point on we re-introduce the superscript i as we will be comparing the attack estimates for all $i \in \mathcal{I}$.

IV. Attack Identification and State Reconstruction

As shown in section III, if $i \in \mathcal{I}_\mathcal{A}$ the state and attack are correctly estimated. This means we can perform secure state estimation if $\mathcal{I}_\mathcal{A}$ is identified. We propose to incrementally build an estimate $\hat{\mathcal{I}}_\mathcal{A}$ of $\mathcal{I}_\mathcal{A}$ via the rule: add i to $\hat{\mathcal{I}}_\mathcal{A}$ if

$$f^i_{\text{est}}(\mathbf{0}, \hat{f}^i_2, \mathbf{0}) = f^i_{\text{est}}(f^i_1[0:t], f^i_2, f^i_3[0:t]), \tag{11}$$

Note that the right-hand side of (11) is obtained from ν in (7) and the left-hand side is calculated using (9) where after initial convergence we have $e_1 = 0$. First, we prove that all $i \in \mathcal{I}_\mathcal{A}$ will be found:

Lemma 4: All observers that provide a correct state and attack estimate will be identified, i.e. $\mathcal{I}_\mathcal{A} \subseteq \hat{\mathcal{I}}_\mathcal{A}$.

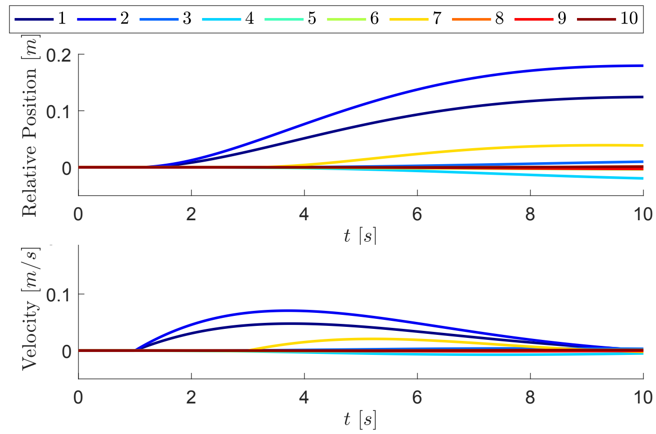

Fig. 1. Position and velocity of all subsystems. Colours correspond to the different subsystems.

Proof: For any $i \in \hat{\mathcal{I}}_\mathcal{A}$ we have $f^i_3[0:t] = f^i_1[0:t] = \mathbf{0}$ and $\hat{f}^i_2 = f^i_2$. Therefore, f^j_{est} will have the same input on both sides such that equation (11) holds. ■

The test in (11) might however identify an $i \notin \mathcal{I}_\mathcal{A}$ if the attack is defined as a zero-dynamics attack on f^i_{est}. Therefore we introduce the following lemma.

Lemma 5: f^i_{est} has no zero-dynamics for all i and for all possible sets of attacked subsystems $\mathcal{A}$.

Proof: The proof is presented in the appendix. ■

Theorem 1: All and only the observers providing a correct state and attack estimate will be identified, i.e. $\hat{\mathcal{I}}_\mathcal{A} = \mathcal{I}_\mathcal{A}$.

Proof: f^i_{est} has no zero-dynamics by Lemma 5. Therefore, (11) holds only if $f^i_1[0:t] = f^i_3[0:t] = \mathbf{0}$, which is equivalent to $i \in \mathcal{I}_\mathcal{A}$. With Lemma 4 this means that (11) holds iff $i \in \mathcal{I}_\mathcal{A}$, which is equivalent to $\hat{\mathcal{I}}_\mathcal{A} = \mathcal{I}_\mathcal{A}$. ■

Any observer $i \in \hat{\mathcal{I}}_\mathcal{A}$ can now be used to perform secure state estimation. Furthermore, the attacked subsystems can be identified from $\hat{\mathcal{I}}_\mathcal{A}$. To this end denote $\mathcal{A}_i$ as set $\mathcal{A}$ according to hypothesis i. Then we can identify the attacks as

$$\hat{\mathcal{A}} = \mathcal{N} \setminus \bigcup_{i \in \hat{\mathcal{I}}_\mathcal{A}} (\mathcal{N} \setminus \mathcal{A}_i), \tag{12}$$

where $\hat{\mathcal{A}} = \mathcal{A}$ if $\hat{\mathcal{I}}_\mathcal{A} = \mathcal{I}_\mathcal{A}$.

V. Numerical Results

The approach is verified on the mass-spring-damper system from [22] which represents subsystems as in (1) with

$$\begin{aligned} A_1 &= A_N = \begin{bmatrix} 0 & 1 \\ -0.3 & -0.1 \end{bmatrix}, \\ A_\ell &= \begin{bmatrix} 0 & 1 \\ -0.4 & -0.1 \end{bmatrix} \text{ for } \ell \in \{2, \dots, N-1\}, \\ A_{\ell j} &= \begin{bmatrix} 0 & 0 \\ 0.1 & 0 \end{bmatrix} \text{ for } \ell \in \mathcal{N}, j \in \{\ell - 1, \ell + 1\}, \\ C_\ell &= [1 \quad 0] \text{ for } \ell \in \mathcal{N} \;;\; B_\ell = \begin{bmatrix} 0 \\ 0.2 \end{bmatrix} \text{ for } \ell \in \mathcal{N}, \end{aligned} \tag{13}$$

and all other matrices are zero. We connect $N = 10$ of such subsystems to obtain a model as in (1) and applied a stabilizing control law to this system.

This system has 10 measurements and each subsystem has 2 potential attacks, one on the actuator and one on

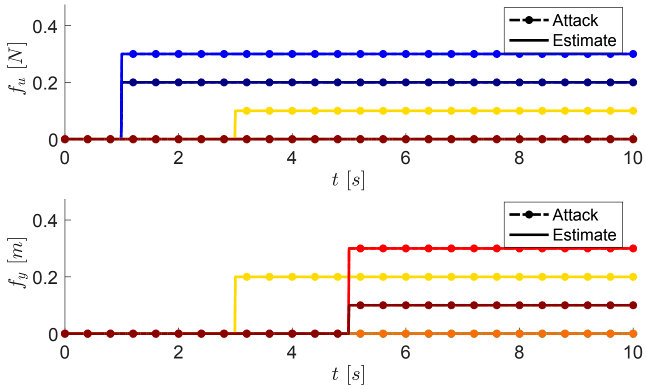

Fig. 2. Performed attacks and attack estimate of the observer that is identified to be correct. Colours correspond to subsystems as in figure 1. The attack estimate line overlays the attack line.

the measurement. Therefore at most 5 subsystems can be attacked while still satisfying Proposition 1. In this section, we consider attacked subsystems $\mathcal{A} = \{1, 2, 7, 9, 10\}$ on which the attacks as shown in Figure 2 are performed. Note that no attacks are present during the first 1 s. The state response to these attacks is shown in Figure 1. Note that the colours in Figure 2 correspond to the legend in Figure 1.

To identify the attack and perform secure state estimation we designed $|\mathcal{I}| = 252 = \binom{10}{5}$ SMOs based on all hypotheses of the attacked subsystems. For the considered set off attacked subsystems $\mathcal{A} = \{1, 2, 7, 9, 10\}$ we have that only hypothesis 55 is correct, i.e. $\mathcal{I}_{\mathcal{A}} = \{55\}$.

For each hypothesis in $\mathcal{I}$, the test in (11) was performed. The result of this test over time is shown in Figure 3. One can see that initially, for $t \in [0,\, 1]$, all hypotheses are identified as correct, i.e. $\hat{\mathcal{I}}_{\mathcal{A}} = \mathcal{I}$. However, with every attack that becomes active, a number of hypotheses are rejected. After $t = 5$ s only the correct hypothesis $\hat{\mathcal{I}}_{\mathcal{A}} = \mathcal{I}_{\mathcal{A}} = \{55\}$ is identified as correct. In figure 4 it is shown that the attack can be correctly identified based on $\hat{\mathcal{I}}_{\mathcal{A}}$ using (12).

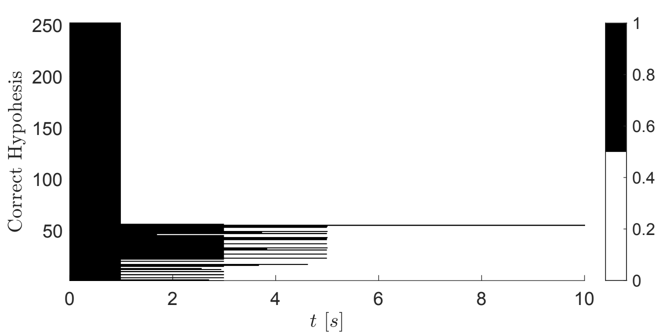

Fig. 3. Black lines indicate the members of set $\hat{\mathcal{I}}_{\mathcal{A}}$ over time. After $t = 5$ s it holds $\hat{\mathcal{I}}_{\mathcal{A}} = \{55\}$

In Figure 2 the attack estimate produced by the observer corresponding to the correct hypothesis $i = 55$ is shown. One can see that, as expected, the fault estimation is very small. The same holds for the state estimation error. The maximum

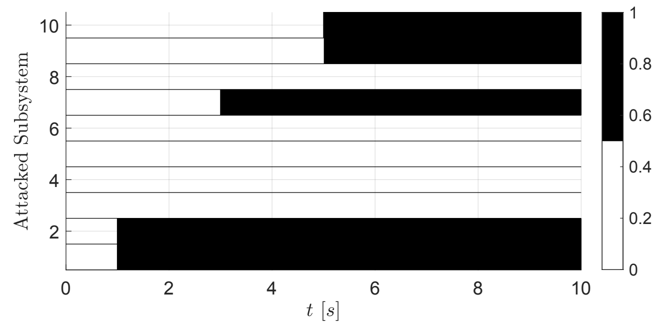

Fig. 4. Solid black color indicates the subsystems that are identified to be attacked over time.

and average mean-squared-errors over all state estimates are $2.4 \cdot 10^{-3}$ and $2.1 \cdot 10^{-6}$ respectively.

VI. Conclusion

Attacks in interconnected systems can fully compromise an unknown, but limited, part of its subsystems. Existing research on secure state estimation (SSE), however, mainly addresses attacks affecting only measurements. In this paper, a sliding mode observer (SMO)-based method to SSE has been proposed that can address attacks that fully compromise subsystems, i.e. affect their actuator and measurements. Furthermore, the proposed method allows for as many attacks as measurements. This represents a much larger class of attacks than SSE is currently available for.

The SMO-based SSE uses a bank of SMOs of which it is proven that at least one can provide correct state and attack estimates. The capability of the SMOs to estimate the attacks has been used to identify which SMOs provide correct state estimates. Specifically, the direct relation between attack and attack estimate has been derived, which is then used to identify the observers that produce correct attack estimates. We prove that using this method we can identify the attacked subsystems. The method has been demonstrated on a system of 10 interconnected mass-spring-damper subsystems.

This paper presents a proof of concept of SMO-based SSE for attacks that fully compromise subsystems. There are, however, many interesting venues for future work to expand on this concept. Firstly, it is interesting to develop this approach for non-linear plants or considering model and measurement uncertainty. Secondly, one might look at distributed implementations of the scheme.

Appendix

Proof: **(Lemma 5)**
Following the approach in [23] we write an equivalent system for each combination of attacked subsystems $\mathcal{A}$ as

$$\begin{cases} \dot{x} = Ax + \mathbf{B}\mathbf{f}_u\,, \\ y = Cx + \mathbf{D}\mathbf{f}_y\,, \end{cases}$$

where $\mathbf{B}$ and $\mathbf{D}$ are the columns of B and D corresponding to non-zero entries of f_u and f_y, denoted by $\mathbf{f}_u$ and $\mathbf{f}_y$. As, by Proposition 1, at most p attacks are active we can use the Rosenbrock matrix to prove f_{est} has no zero-dynamics.

Assumption 2 states that the full system is input observable, which is equivalent to not having zero dynamics.[19] Therefore $\nexists\ s$ for which

$$\begin{bmatrix} sI - A & -\mathbf{B} & 0 \\ C & 0 & \mathbf{D} \end{bmatrix} \tag{14}$$

loses rank. Below we will go through all derivation steps in Section III and prove they cannot cause zero dynamics.

First, for all $i \in \mathcal{I}$ we perform a lossless transformation, which leads to a Rosenbrock matrix with equivalent zero-dynamics to (14). Then in Section III-A we filter part of the measurements and obtain Rosenbrock matrix

$$\begin{bmatrix} sI - A_{11} & -A_{12} & 0 & -\mathbf{B}_{11} & 0 & 0 & 0 \\ -A_{21} & sI - A_{22} & 0 & 0 & -\mathbf{B}_{22} & 0 & 0 \\ A_f C_{21} & A_f C_{22} & sI - A_f & 0 & 0 & A_f \mathbf{D}_{22} & 0 \\ C_{11} & C_{12} & 0 & 0 & 0 & 0 & \mathbf{D}_{11} \\ 0 & 0 & I & 0 & 0 & 0 & 0 \end{bmatrix}$$

where the same boldface notation is used for B and D as in (14). Here row 5 and column 3 can be removed without affecting the zero dynamics. Furthermore, row 3 can be pre-multiplied with A_f^{-1}, which is a lossless transformation as A_f is full rank. The resulting matrix is the Rosenbrock matrix of the system in (3).

In Section III-B a row is added to the system output, which cannot introduce zero-dynamics. Then in Section III-C the SMO standard form in (6) is obtained by a lossless transformation not affecting the zero-dynamics. From the Rosenbrock matrix of the system in (6) we provide equivalence to the Rosenbrock matrix of f_{est} in three steps as

$$\begin{bmatrix} sI - A_1 & -A_2 & -\mathbf{B}_1 & 0 & 0 \\ -A_3 & sI - A_4 & -\mathbf{B}_{21} & -\mathbf{B}_2 & 0 \\ 0 & I & 0 & 0 & \mathbf{D} \end{bmatrix}$$
$$\Updownarrow 1)$$
$$\begin{bmatrix} sI - A_1 & -A_2\mathbf{D} & -\mathbf{B}_1 & 0 \\ -A_3 & (sI - A_4)\tilde{D} & -\mathbf{B}_{21} & -\mathbf{B}_2 \end{bmatrix}$$
$$\Updownarrow 2)$$
$$\begin{bmatrix} sI - A_1 & -A_2\mathbf{D} & -\mathbf{B}_1 & 0 \\ -B_2^{\dagger} A_3 & B_2^{\dagger}(sI - A_4)\tilde{D} & -B_2^{\dagger}\mathbf{B}_{21} & -B_2^{\dagger}\mathbf{B}_2 \\ -T_B A_3 & T_B(sI - A_4)\tilde{D} & -T_B\mathbf{B}_{21} & -T_B\mathbf{B}_2 \end{bmatrix} \tag{15}$$
$$\Updownarrow 3)$$
$$\begin{bmatrix} sI - A_1 & -A_2\mathbf{D} & \mathbf{B}_1 & 0 & 0 \\ 0 & sI & 0 & 0 & -I \\ -B_2^{\dagger} A_3 & -B_2^{\dagger} A_4 \tilde{D} & B_2^{\dagger}\mathbf{B}_{21} & B_2^{\dagger}\mathbf{B}_2 & B_2^{\dagger}\mathbf{D} \\ -T_B A_3 & -T_B A_4 \tilde{D} & T_B\mathbf{B}_{21} & T_B\mathbf{B}_2 & T_B\mathbf{D} \end{bmatrix}$$

Here the steps taken in each transformation are listed below.

1) Subtract second column $\mathbf{D}$ times from fifth column; Remove second column and third row; Multiply fourth column with $-I$; Move fourth column between columns 1 and 2.
2) Pre-multiply second row by full rank matrix $\begin{bmatrix} B_2^{\dagger} \\ T_B \end{bmatrix}$.
3) From bottom to top: Add second row $B_2^{\dagger}\mathbf{D}$ times to third row and $T_B\mathbf{D}$ times to fourth row; Remove second row and fifth column; multiply third and fourth columns with $-I$.

The last matrix in (15) is the Rosenbrock matrix of the system that defines f_{est}. ■

References

[1] A. A. Cárdenas, S. Amin, and S. Sastry, "Research challenges for the security of control systems," in *HOTSEC*, 2008.

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

[3] L. Lamport, R. Shostak, and M. Pease, "The byzantine generals problem," *ACM Trans. Program. Lang. Syst.*, vol. 4, no. 3, p. 382–401, jul 1982.

[4] M. S. Chong, M. Wakaiki, and J. P. Hespanha, "Observability of linear systems under adversarial attacks," in *2015 American Control Conference (ACC)*, 2015, pp. 2439–2444.

[5] Y. Shoukry, P. Nuzzo, A. Puggelli, A. L. Sangiovanni-Vincentelli, S. A. Seshia, and P. Tabuada, "Secure state estimation for cyber-physical systems under sensor attacks: A satisfiability modulo theory approach," *IEEE T. Autom. Contr.*, vol. 62, no. 10, pp. 4917–4932, 2017.

[6] J. Kim, C. Lee, H. Shim, Y. Eun, and J. H. Seo, "Detection of sensor attack and resilient state estimation for uniformly observable nonlinear systems having redundant sensors," *IEEE Trans. on Automatic Control*, vol. 64, no. 3, pp. 1162–1169, 2019.

[7] X. He, X. Ren, H. Sandberg, and K. H. Johansson, "How to secure distributed filters under sensor attacks," *IEEE Trans. on Automatic Control*, vol. 67, no. 6, pp. 2843–2856, 2022.

[8] M. U. B. Niazi, A. Alanwar, M. S. Chong, and K. H. Johansson, "Resilient set-based state estimation for linear time-invariant systems using zonotopes," *arXiv preprint arXiv:2211.08474*, 2022.

[9] Y. Mao, A. Mitra, S. Sundaram, and P. Tabuada, "On the computational complexity of the secure state-reconstruction problem," *Automatica*, vol. 136, p. 110083, 2022.

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

[11] H. Fawzi, P. Tabuada, and S. Diggavi, "Security for control systems under sensor and actuator attacks," in *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, 2012, pp. 3412–3417.

[12] ——, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE T. Autom. Contr.*, vol. 59, no. 6, pp. 1454–1467, 2014.

[13] S. Nateghi, Y. Shtessel, C. Edwards, and J.-P. Barbot, "Secure state estimation and attack reconstruction in cyber-physical systems: Sliding mode observer approach," in *Control Theory in Engineering*, C. Volosencu, A. Saghafinia, X. Du, and S. Chakrabarty, Eds. IntechOpen, 2019, ch. 1.

[14] C. P. Tan and C. Edwards, "Sliding mode observers for robust detection and reconstruction of actuator and sensor faults," *Int. J. of Robust and Nonlinear Control*, vol. 13, pp. 443–463, 4 2003.

[15] S. Nateghi, Y. Shtessel, C. Edwards, and J.-P. Barbot, "Resilient control of cyber-physical systems using adaptive super-twisting observer," *Asian J. of Control*, vol. n/a, no. n/a, 2022.

[16] C. Edwards, S. K. Spurgeon, and R. J. Patton, "Sliding mode observers for fault detection and isolation," *Autom.*, vol. 36, pp. 541–553, 2000.

[17] Y. Xiong and M. Saif, "Unknown disturbance inputs estimation based on a state functional observer design," *Automatica*, vol. 39, no. 8, pp. 1389–1398, 2003.

[18] M. Hou and R. Patton, "Input observability and input reconstruction," *Automatica*, vol. 34, no. 6, pp. 789–794, 1998.

[19] H. Shim, J. Back, Y. Eun, G. Park, and J. Kim, *Zero-Dynamics Attack, Variations, and Countermeasures*. Cham: Springer Int. Publishing, 2022, pp. 31–61.

[20] D. V. Efimov and L. Fridman, "A hybrid robust non-homogeneous finite-time differentiator," *IEEE Trans. on Automatic Control*, vol. 56, no. 5, pp. 1213–1219, 2011.

[21] C. Edwards, S. K. Spurgeon, R. J. Patton, and P. Klotzek, "Sliding mode observers for fault detection," *IFAC Proceedings Volumes*, vol. 30, pp. 507–512, 8 1997.

[22] D. Barcelliy, A. Bemporadz, and G. Ripaccioliy, "Hierarchical multi-rate control design for constrained linear systems," in *49th IEEE Conference on Decision and Control (CDC)*, 2010, pp. 5216–5221.

[23] F. Pasqualetti, F. Dorfler, and F. Bullo, "Cyber-physical security via geometric control: Distributed monitoring and malicious attacks," in *Conference on Decision and Control*, 2012, pp. 3418–3425.