2022 American Control Conference (ACC)
Atlanta, USA, June 8-10, 2022

Resilient Interval Observer for Simultaneous Estimation of States, Modes and Attack Policies

Mohammad Khajenejad Zeyuan Jin Sze Zheng Yong

Abstract— **This paper considers the problem of designing interval observers for hidden mode switched nonlinear systems with bounded noise signals that are compromised by false data injection and switching attacks. The proposed observer consists of three components: i) a bank of mode-matched observers, which simultaneously estimates the corresponding mode-matched continuous states and discrete states (modes), as well as learns a model of the unknown attack policy, ii) a mode observer that eliminates the incompatible modes based on a residual-based set-membership criterion, and iii) a global fusion observer that combines the outputs of i) and ii). Moreover, in addition to showing the correctness, stability and convergence of the mode-matched estimates, we provide sufficient conditions to guarantee that all false modes will be eliminated after sufficiently large finite time steps, i.e., the system is mode-detectable under the proposed observer.**

I. INTRODUCTION

Computation and communication constituents are tightly intertwined in Cyber-Physical Systems (CPS). While this coupling can enhance the functionality of control systems and improve their performance, it might also become a source of vulnerability to faults or attacks. On the other hand, given various sources of real world uncertainties, complete information/direct knowledge of the decisions and intentions of other systems/agents, is not available to autonomous decision makers, e.g., self-driving cars or robots. These safety-critical systems can be studied using a general framework of *hidden mode hybrid/switched systems* (HMHS, see, e.g., [1] and references therein). The ability to estimate the continuous states, attacks/unknown inputs and modes/discrete states of such systems is important for monitoring them as well as for designing safe and secure (optimal) feedback controllers.

There has been a relatively large body of literature on the problem of designing filters/observers for hidden mode systems without considering unknown inputs/faults/data injection attacks, e.g., in [2] and references therein. For a stochastic setting, extensions were proposed, e.g., in [1], to obtain state and unknown input *point* estimates, i.e., the most likely or best single estimates. However, especially when hard guarantees or bounds are important, it might be preferable to consider *set-valued* uncertainties, e.g., bounded-norm noise. Moreover, probabilistic distributions/stochastic characteristics of uncertainty are often unavailable in real world applications. Consequently, to estimate the *set* of compatible states, *set-valued* or *set-membership* observers, e.g., [3], have been proposed. Later, the study in [4] extended this framework to include estimation of unknown inputs. Nonetheless, these approaches are not directly applicable to systems with hidden modes that are considered in this paper.

A common approach to consider hidden modes for representing attack or fault models is to construct *residual* signals (see, e.g., [1], [2], [5]), where to distinguish between consistent and inconsistent modes, some residual-based criteria/thresholds are used. The work in [6] presented a robust control-inspired approach for linear systems with bounded-norm noise that consists of local estimators, residual detectors, and a global fusion detector for resilient state estimation against sparse data injection attacks. Similar residual-based approaches have been proposed for uniformly observable nonlinear systems in [7] and some classes of nonlinear systems in [8], where only sensors were compromised by sparse attacks, which is a special case of hidden mode switched systems discussed in our previous works [1], [9], where actuators were also compromised by attack signals.

On the other hand, when the system model is not exactly known, in order to find *a set of dynamics* that *frame/bracket* the unknown system dynamics [10], set-valued data-driven approaches have been developed to use input-output data to *abstract* or over-approximate unknown dynamics or functions [10], [11], under the assumption that the unknown dynamics is continuous, e.g., [11]. In our previous work [12], we leveraged interval observers for such data-driven models, for resilient state and data injection attack estimation, assuming that the attack signal has an unknown dynamics. In this work, we assume mode/switching attacks in addition to data injection attacks, where the attack signals are governed by an unknown and to-be-learned *attack policy*.

To tackle this problem, leveraging a multiple-model framework proposed in our previous works [9], [13], we first design a bank of mode-matched set-valued observers, where we combine a model-based interval observer approach used in [12], [14], with our previously introduced set-membership learning technique [15], to derive set-valued mode-matched estimates for the states and attack signal values, as well as to learn model abstractions/over-approximations for the attack policy, where we derive several desired properties for the mode-matched estimates, such as *correctness*, *stability* and *convergence*. Then, we introduce a novel *elimination-based* mode observer, based on a set-membership criterion, to eliminate inconsistent modes from the bank of observers. Further, we provide sufficient conditions for *mode-detectability*, i.e., all false modes will be eventually ruled out under some reasonable assumptions. Finally, we illustrate the performance of our proposed design through a power system example.

M. Khajenejad is with the Department of Mechanical and Aerospace Engineering, University of California, San Diego, CA, USA (email: mkhajenejad@ucsd.edu). Z. Jin and S.Z. Yong are with the School for Engineering of Matter, Transport and Energy, Arizona State University, Tempe, AZ, USA (e-mail: zjin43,szyong@asu.edu). This entire work was done at ASU and is supported in part by NSF grant CNS-1943545.

978-1-6654-5196-3/$31.00 ©2022 AACC

II. Preliminaries

Notation. $\mathbb{R}^n$, $\mathbb{R}^{n\times m}$ and $\mathbb{D}_n$ denote the n-dimensional Euclidean space, the space of n by m matrices and the set of all diagonal matrices in $\mathbb{R}^{n\times n}$ with their diagonal arguments being 0 or 1. For vectors $v,w\in\mathbb{R}^n$ and a matrix $M\in\mathbb{R}^{p\times q}$, $\|v\|\triangleq\sqrt{v^\top v}$ and $\|M\|$ denote their (induced) 2-norm, and $v\le w$ is an element-wise inequality. The transpose, Moore-Penrose pseudoinverse, (i,j)-th element and rank of M are given by $M^\top$, $M^\dagger$, $M_{i,j}$ and $\mathrm{rk}(M)$, while $M_{(r:s)}$ is a sub-matrix of M, consisting of its r-th through s-th rows, and its row support is $r=\mathrm{rowsupp}(M)\in\mathbb{R}^p$, where $r_i=0$ if the i-th row of M is zero and $r_i=1$ otherwise, $\forall i\in\{1\dots p\}$. Also, $M^+\triangleq\max(M,0_{p\times q})$, $M^-\triangleq M^+-M$ and $|M|\triangleq M^++M^-$. M is a non-negative matrix, if $M_{i,j}\ge 0,\forall(i,j)\in\{1\dots p\}\times\{1\dots q\}$.

Next, we introduce some useful definitions and results that will be used later in our derivations and proofs.

Definition 1 (Interval, Interval Width)**.** *A multi-dimensional interval $\mathcal{I}\subset\mathbb{R}^n$ is the set of all vectors $x\in\mathbb{R}^n$ satisfying $\underline{s}\le x\le\overline{s}$, where $\|\overline{s}-\underline{s}\|$ is called the width of $\mathcal{I}$.*

Proposition 1 (Slight Generalization of [16, Lemma 2])**.** *Let $B\in\mathbb{IR}^{n\times p}$ be an interval matrix satisfying $\underline{B}\le B\le\overline{B}$.*

i) if $A\in\mathbb{R}^{m\times n}$ is a constant matrix, then $A^+\underline{B}-A^-\overline{B}\le AB\le A^+\overline{B}-A^-\underline{B}$.

ii) if $A\in\mathbb{IR}^{m\times n}$ is an interval matrix satisfying $\underline{A}\le A\le\overline{A}$, then $\underline{A}^+\underline{B}^+-\overline{A}^+\underline{B}^--\underline{A}^-\overline{B}^++\overline{A}^-\overline{B}^-\le AB\le\overline{A}^+\overline{B}^+-\underline{A}^+\overline{B}^--\overline{A}^-\underline{B}^++\underline{A}^-\underline{B}^-$

Proof. The results follow from defining x_i as the ith column of B, applying [16, Lemma 2] on A and x_i for all $i\in\mathbb{N}_p$ and then stacking the resulting inequalities. ■

Proposition 2 (Parallel Affine Abstractions [12])**.** *Let the entire space be defined as $\mathbb{X}$ and suppose that $\mathbb{X}$ is bounded. Consider the vector fields $\overline{\psi}(.),\underline{\psi}(.):\mathbb{X}\subset\mathbb{R}^{n'}\to\mathbb{R}^{m'}$ satisfying $\underline{\psi}(x)\le\overline{\psi}(x),\forall x\in\mathbb{X}$, a (given) global parallel affine abstraction with known $(\mathbb{A}^\psi,\overline{e}^\psi,\underline{e}^\psi)$ on $\mathbb{X}$, i.e.,*

$$\mathbb{A}^\psi x+\underline{e}^\psi\le\underline{\psi}(x)\le\overline{\psi}(x)\le\mathbb{A}^\psi x+\overline{e}^\psi,\forall x\in\mathbb{X}.\quad(1)$$

and the following Linear Program (LP):

$$\min_{\theta^\psi_\mathcal{B},A^\psi_\mathcal{B},\overline{e}^\psi_\mathcal{B},\underline{e}^\psi_\mathcal{B}}\theta^\psi_\mathcal{B}\quad(2a)$$

$$\begin{aligned}s.t.\ & A^\psi_\mathcal{B}x_s+\underline{e}^\psi_\mathcal{B}+\sigma^\psi\le\underline{\psi}(x_s)\le\overline{\psi}(x_s)\le A^\psi_\mathcal{B}x_s+\overline{e}^\psi_\mathcal{B}-\sigma^\psi,\\ & \overline{e}^\psi_\mathcal{B}-\underline{e}^\psi_\mathcal{B}-2\sigma^\psi\le\theta^\psi\mathbf{1}_{m'},\\ & \underline{e}^\psi-\underline{e}^\psi_\mathcal{B}\le(A^\psi_\mathcal{B}-\mathbb{A}^\psi)x_s\le\overline{e}^\psi-\overline{e}^\psi_\mathcal{B},\forall x_s\in\mathcal{V}_\mathcal{B},\end{aligned}\quad(2b)$$

where $\mathcal{B}=[\underline{x},\overline{x}]\subseteq\mathbb{X}$ is a local interval domain and $\mathcal{V}_\mathcal{B}$ being its maximal, minimal and set of vertices, respectively, $\mathbf{1}_m\in\mathbb{R}^m$ is a vector of ones, σ^ψ is given in [17, Proposition 1 and (8)] for different classes of continuous vector fields. Then, $(A^\psi_\mathcal{B},\overline{e}^\psi_\mathcal{B},\underline{e}^\psi_\mathcal{B})$ are the local parallel affine abstraction matrices for the pair of functions $\overline{\psi}(.),\underline{\psi}(.)$ on $\mathcal{B}$, i.e.,

$$A^\psi_\mathcal{B}x+\underline{e}^\psi_\mathcal{B}\le\underline{\psi}(x)\le\overline{\psi}(x)\le A^\psi_\mathcal{B}x+\overline{e}^\psi_\mathcal{B},\forall x\in\mathcal{B}.\quad(3)$$

Definition 2 (Mixed-Monotone Mappings and Decomposition Functions)**.** *[18, Definition 4] A mapping $f:\mathcal{X}\subseteq\mathbb{R}^n\to\mathcal{T}\subseteq\mathbb{R}^m$ is mixed-monotone if there exists a decomposition function $f_d:\mathcal{X}\times\mathcal{X}\to\mathcal{T}$ satisfying: i) $f_d(x,x)=f(x)$, ii) $x_1\ge x_2\Rightarrow f_d(x_1,y)\ge f_d(x_2,y)$ and iii) $y_1\ge y_2\Rightarrow f_d(x,y_1)\le f_d(x,y_2)$.*

Proposition 3. *[19, Theorem 1] Let $f:\mathcal{X}\subseteq\mathbb{R}^n\to\mathcal{T}\subseteq\mathbb{R}^m$ be a mixed-monotone mapping with decomposition function $f_d:\mathcal{X}\times\mathcal{X}\to\mathcal{T}$ and $\underline{x}\le x\le\overline{x}$, where $\underline{x},x,\overline{x}\in\mathcal{X}$. Then $f_d(\underline{x},\overline{x})\le f(x)\le f_d(\overline{x},\underline{x})$.*

Corollary 1 (Nonlinear Bounding)**.** *Let $f:\mathcal{X}\subseteq\mathbb{R}^n\to\mathcal{T}\subseteq\mathbb{R}^m$ satisfies the assumptions in Propositions 2 and 3. Then, for all $\underline{x},x,\overline{x}\in\mathcal{X}$ satisfying $\underline{x}\le x\le\overline{x}$, the following inequality holds: $\underline{f}(\overline{x},\underline{x})\le f(x)\le\overline{f}(\overline{x},\underline{x})$, where*

$$\begin{aligned}\overline{f}(\overline{x},\underline{x})&=\min(f_d(\overline{x},\underline{x}),A^{f+}\overline{x}-A^{f-}\underline{x}+\overline{e}^f),\\ \underline{f}(\overline{x},\underline{x})&=\max(f_d(\underline{x},\overline{x}),A^{f+}\underline{x}-A^{f-}\overline{x}+\underline{e}^f),\end{aligned}\quad(4)$$

f_d is a decomposition function of f (cf. Definition 2) and $A^f,\overline{e}^f,\underline{e}^f$ are the affine abstraction slope and errors of f, computed over the interval $[\underline{x},\overline{v}]$, through Proposition 2.

Proof. The results directly follow from Propositions 1–3. ■

Note that the decomposition function of a vector field is not unique and a specific one is given in [18, Theorem 2]: If a vector field $q=\begin{bmatrix}q_1^\top\dots q_n^\top\end{bmatrix}^\top:X\subseteq\mathbb{R}^n\to\mathbb{R}^m$ is differentiable and its partial derivatives are bounded with known bounds, i.e., $\frac{\partial q_i}{\partial x_j}\in(a^q_{i,j},b^q_{i,j}),\forall x\in X\in\mathbb{R}^n$, where $a^q_{i,j},b^q_{i,j}\in\mathbb{R}$, then q is mixed-monotone with a decomposition function $q_d=\begin{bmatrix}q_{d1}^\top\dots q_{di}^\top\dots q_{dn}^\top\end{bmatrix}^\top$, where $q_{di}(x,y)=q_i(z)+(\alpha^q_i-\beta^q_i)^\top(x-y),\forall i\in\{1,\dots,n\}$, and $z,\alpha^q_i,\beta^q_i\in\mathbb{R}^n$ can be computed in terms of $x,y,a^q_{i,j},b^q_{i,j}$ as given in [18, (10)–(13)]. Consequently, for $x=[x_1\dots x_j\dots x_n]^\top$, $y=[y_1\dots y_j\dots y_n]^\top$, we have

$$q_d(x,y)=q(z)+C^q(x-y),\quad(5)$$

where $C^q\triangleq\begin{bmatrix}[\alpha^q_1-\beta^q_1]\dots[\alpha^q_i-\beta^q_i]\dots[\alpha^q_m-\beta^q_m]\end{bmatrix}^\top\in\mathbb{R}^{m\times n}$, with α^q_i,β^q_i given in [18, (10)–(13)], $z=[z_1\dots z_j\dots z_m]^\top$ and $z_j=x_j$ or y_j (dependent on the case, cf. [18, Theorem 1 and (10)–(13)] for details). More recently, a tractable way for computing tight remainder-form decomposition functions was proposed in [20].

III. Problem Statement

System Assumptions. Consider a discrete-time hidden mode switched nonlinear system with bounded-norm noise and unknown inputs (i.e., a hybrid system with nonlinear and noisy system dynamics in each mode, and the mode and some inputs are not known/measured):

$$\begin{aligned}x_{k+1}&=\hat{f}^q(x_k,u^q_k,G^qd^q_k,w_k)\triangleq f^q(x_k,d^q_k,w_k),\\ y_k&=\hat{g}^q(x_k,u^q_k,H^qd^q_k,v_k)\triangleq g^q(x_k,d^q_k,v_k),\\ d^q_k&=\hat{\mu}^q(x_k,u^q_k)\triangleq\mu^q(x_k),\end{aligned}\quad(6)$$

where $x_k\in\mathbb{R}^n$ is the continuous system state and $q\in\mathcal{Q}=\{1,2,\dots,Q\}$ is the hidden discrete state or *mode*. For each (fixed) mode q, $u^q_k\in U^q_k\subset\mathbb{R}^m$ is the *known* input, $d^q_k\in\mathbb{R}^p$ is the unknown but *sparse* input, i.e., every vector d^q_k has precisely $\rho\in\mathbb{N}$ nonzero elements where ρ is a known parameter and $y_k\in\mathbb{R}^l$ is the measured output. The unknown input signal d_k^q is considered as the realization of an attacker's *unknown policy* $\mu^q : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^s \to \mathbb{R}^p$, which is an unknown mapping from state and known input to the set of attack signals. Moreover, $w_k \in \mathcal{W} \triangleq [\underline{w}, \overline{w}] \subset \mathbb{R}^{n_w}$ and $v_k \in \mathcal{V} \triangleq [\underline{v}, \overline{v}] \subset \mathbb{R}^{n_v}$ are bounded process and measurement disturbances with known minimal and maximal values $\underline{w}, \overline{w}, \underline{v}, \overline{v}$, respectively. Further, the mappings f, g, as well as the matrices $G^q \in \mathbb{R}^{n \times p}$ and $H^q \in \mathbb{R}^{l \times p}$ are known.

More precisely, G^q and H^q represent the different hypothesis for each mode $q \in \mathcal{Q}$, about the sparsity pattern of the unknown inputs, which in the context of sparse attacks corresponds to which actuators and sensors are attacked or not attacked. In other words, we assume that $G^q = G\mathbb{I}_G^q$ and $H^q = H\mathbb{I}_H^q$ for some input matrices $G \in \mathbb{R}^{n \times t_a}$ and $H \in \mathbb{R}^{l \times t_s}$, where t_a and t_s are the number of vulnerable actuator and sensor signals respectively. Note that $\rho_a^q \leq t_a \leq m$ and $\rho_s^q \leq t_s \leq l$, where ρ_a^q (ρ_s^q) is the number of attacked actuator (sensor) signals and clearly cannot exceed the number of vulnerable actuator (sensor) signals, which in turn cannot exceed the total number of actuators (sensors). Further, we assume that the maximum number of unknown inputs/attacks in each mode is known and equals $\rho = \rho_a + \rho_s$ (sparsity assumption). Moreover, the *index matrix* $\mathbb{I}_G^q \in \mathbb{R}^{t_a \times \rho}$ ($\mathbb{I}_H^q \in \mathbb{R}^{t_s \times \rho}$) represents the sub-vector of $d_k \in \mathbb{R}^\rho$ that indicates signal magnitude attacks on the actuators (sensors).

We are interested in estimating the state trajectories, as well as the unknown mode and the attack policy mapping in the system in (6), when they are initialized in a given interval $\mathcal{X}_0 \subset \mathcal{X} \subset \mathbb{R}^n$. Furthermore, we assume the following:

Assumption 1. *The vector fields f, g are known and Lipschitz continuous (hence, mixed-monotone). Moreover, the input u_k^q and output y_k signals are known at all times and for all modes. The set of all possible modes, $\mathcal{Q}$, is also known.*

Assumption 2. *Given mode q, the attacker's policy mapping $\mu^q(\cdot) = [\mu_1^{q\top}(\cdot), \ldots, \mu_p^{q\top}(\cdot)]^\top$ is unknown, but each $\mu_j^q(\cdot), \forall j \in \{1, \ldots, p\}$ is known to be Lipschitz continuous. Moreover, for simplicity and without loss of generality we assume that the Lipschitz constants $L_j^{\mu^q}, \forall j \in \{1, \ldots, p\}$ are known, otherwise, they can be estimated with any desired precision using the approach in [15, Equation (12) and Proposition 3].*

Assumption 3. *There is only one "true" mode, i.e., the true mode $q^* \in \mathcal{Q}$ is constant over time.*

Note that the approach in our paper can be easily extended to handle mode-dependent f, g, $\overline{w}$, $\underline{w}$, $\overline{v}$ and $\underline{v}$, but is omitted to simplify the notations. Further, we define the notions of *framers*, *correctness* and *stability*, used throughout the paper.

Definition 3 (Correct Interval Framers)**.** *Given a hidden mode switched nonlinear system* (6)*, let us define the augmented state $z_k \triangleq [x_k^\top \; d_k^\top]^\top$, for all $k \in \mathbb{K} \triangleq \mathbb{N} \cup \{0\}$, where $d_k \triangleq d_k^{q^*}$ is the true attack signal. The sequences $\{\overline{z}_k, \underline{z}_k\}_{k=0}^\infty$ are called upper and lower framers for the augmented states of system* (6)*, if $\forall k \in \mathbb{K}$, $\underline{z}_k \leq z_k \leq \overline{z}_k$. In other words, starting from the initial interval $z_0 \in [\underline{z}_0, \overline{z}_0]$, the true augmented state of the system in* (6)*, z_k, is guaranteed to evolve within the interval flow-pipe $[\underline{z}_k, \overline{z}_k]$, for all $k \in \mathbb{K}$. Finally, any algorithm that returns framers for the states of system* (6) *is called a* correct *interval framer.*

Definition 4 (Stability)**.** *The mode-matched observer* (8a)–(11b) *is stable, if the sequence of interval widths $\{\|\Delta_k^{z^q}\| \triangleq \|\overline{z}_k^q - \underline{z}_k^q\|\}_{k=0}^\infty$ is uniformly bounded, and consequently, the sequence of estimation errors $\{\|\tilde{z}_k^q\| \triangleq \max(\|z_k^q - \underline{z}_k^q\|, \|\overline{z}_k^q - z_k^q\|)$ is also uniformly bounded.*

An interval observer is then an estimator that is both correct and stable. Using the modeling framework above, the simultaneous state, hidden mode and policy estimation problem is threefold and can be stated as follows:

Problem 1. *Given a discrete-time bounded-error hidden mode switched nonlinear system with unknown inputs* (6) *and assuming that Assumptions 1–3 hold,*

i) Design a bank of mode-matched observers that for each mode, conditioned on the mode being the true mode, finds uniformly bounded set estimates of compatible (augmented) states and learns a guaranteed model abstraction of the attacker's policy.

ii) Develop a mode observer via elimination and the corresponding criteria to eliminate false modes.

iii) Find sufficient conditions for eliminating all false modes.

IV. Proposed Observer Design

Leveraging a multiple-model approach similar to [9], [13], our goal in this section is to propose an observer for simultaneous mode, state and attack policy (SMSP) estimation, i.e., to find set estimates $\hat{\mathcal{X}}_k$, $\hat{\mathcal{D}}_k$ and $\hat{\mathcal{Q}}_k$ for the states x_k, attacks d_k and modes $q \in \mathcal{Q}$ at time step k, respectively, as well as to compute a model abstraction $\{\overline{\mu}_k, \underline{\mu}_k\}_{k \in \mathbb{K}}$ for the attack policy, such that $\underline{\mu}_k(x_k) \leq \mu(x) \leq \overline{\mu}(x_k)$ for all $k \in \mathbb{K}$.

A. Multiple-Model Approach: An Overview

Similar to the approach in [13], we propose a three-step multiple-model design consisting of: (i) a bank of mode-matched interval observers to obtain mode-matched state and attack estimates, as well as mode-matched policy abstractions/over-approximations, (ii) a mode estimation algorithm to eliminate incompatible modes using residual detectors, and (iii) a global fusion observer that outputs the desired set-valued mode, attack (policy) and state estimates.

1) Mode-Matched Set-Valued State and Attack Policy Observer: First, we design a bank of mode-matched observers, which consists of $Q \triangleq |\mathcal{Q}|$ simultaneous state, attack and policy mode-matched interval observers, designed in a similar manner as our approach in [12], with the difference that in [12], the unknown input (i.e., attack) signal is treated as a state with unknown and to-be-learned dynamics, whereas in the current work, the attack signal is governed by an unknown policy/state feedback law, i.e., an unknown function of the actual state, that should be learned/approximated. With that in mind, given mode q, each mode-matched interval observer at time step $k \in \mathbb{N}$, recursively returns

$$\hat{\mathcal{X}}_k^q \triangleq [\underline{x}_k^q, \overline{x}_k^q], \; \hat{\mathcal{D}}_k^q \triangleq [\underline{d}_k^q, \overline{d}_k^q], \; \{\underline{\mu}_k^q, \overline{\mu}_k^q\}, \quad (7)$$

such that the true state x_k and unknown input d_k are contained in $\hat{\mathcal{X}}_k^q$ and $\hat{\mathcal{D}}_k^q$, respectively, i.e., $x_k \in \hat{\mathcal{X}}_k^q$, $d_k \in \hat{\mathcal{D}}_k^q$ and the true attack policy $d_k = \mu(x_k)$ is also contained in the learned function framer, i.e., $\mu(x_k) \in [\underline{\mu}_k^q(x_k), \overline{\mu}_k^q(x_k)]$.

This can be achieved through the following steps (with known $\underline{x}_0$ and $\overline{x}_0$ such that $\underline{x}_0 \leq x_0 \leq \overline{x}_0$), where we defined the augmented state $z_k^q \triangleq \begin{bmatrix} x_k^\top & d_k^{q\top} \end{bmatrix}^\top$, propagated framers $\overline{z}_k^{q,p} = \begin{bmatrix} \overline{x}_k^{q,p\top} & \overline{d}_k^{q,p\top} \end{bmatrix}^\top$, $\underline{z}_k^{q,p} = \begin{bmatrix} \underline{x}_k^{q,p\top} & \underline{d}_k^{q,p\top} \end{bmatrix}^\top$ and updated framers $\overline{z}_k^q = \begin{bmatrix} \overline{x}_k^{q\top} & \overline{d}_k^{q\top} \end{bmatrix}^\top$, $\underline{z}_k^q = \begin{bmatrix} \underline{x}_k^{q\top} & \underline{d}_k^{q\top} \end{bmatrix}^\top$:

State Propagation (with $\overline{f}^q, \underline{f}^q$ in Corollary 1):

$$\begin{bmatrix} \overline{x}_k^{q,p} \\ \underline{x}_k^{q,p} \end{bmatrix} = \begin{bmatrix} \overline{f}^q(\overline{z}_{k-1}^q, \underline{z}_{k-1}^q) \\ \underline{f}^q(\overline{z}_{k-1}^q, \underline{z}_{k-1}^q) \end{bmatrix}, \tag{8a}$$

Attack Policy Learning:

$$\overline{\mu}_{k,j}^q(x_k) = \min_{t \in \{0,\ldots,T-1\}} (\overline{d}_{k-t,j}^q + L_j^{\mu^q} \|x_k - \tilde{x}_{k-t}^q\|) + \varepsilon_{k-t}^{q,j}, \tag{9a}$$

$$\underline{\mu}_{k,j}^q(x_k) = \max_{t \in \{0,\ldots,T-1\}} (\underline{d}_{k-t,j}^q - L_j^{\mu^q} \|x_k - \tilde{x}_{k-t}^q\|) + \varepsilon_{k-t}^{q,j}, \tag{9b}$$

Unknown Input Estimation (with $\overline{\mu}^q, \underline{\mu}^q$ in Corollary 1):

$$\begin{bmatrix} \overline{d}_k^{q,p} \\ \underline{d}_k^{q,p} \end{bmatrix} = \begin{bmatrix} \overline{\mu}^q(\overline{z}_{k-1}^q, \underline{z}_{k-1}^q) \\ \underline{\mu}^q(\overline{z}_{k-1}^q, \underline{z}_{k-1}^q) \end{bmatrix}, \tag{10a}$$

Measurement Update:

$$\begin{bmatrix} \overline{z}_k^q & \underline{z}_k^q \end{bmatrix} = \lim_{i \to \infty} \begin{bmatrix} \overline{z}_{i,k}^{q,u} & \underline{z}_{i,k}^{q,u} \end{bmatrix}, \tag{11a}$$

$$\begin{bmatrix} \overline{x}_k^q & \underline{x}_k^q \\ \overline{d}_k^q & \underline{d}_k^q \end{bmatrix} = \begin{bmatrix} \overline{z}_{k,(1:n)}^q & \underline{z}_{k,(1:n)}^q \\ \overline{z}_{k,(n+1:n+p)}^q & \underline{z}_{k,(n+1:n+p)}^q \end{bmatrix}, \tag{11b}$$

with $j \in \{1 \ldots p\}$, where $\{\tilde{x}_{k-t}^q = \frac{1}{2}(\overline{x}_{k-t}^q + \underline{x}_{k-t}^q)\}_{t=0}^k$ and $\{\overline{d}_{k-t}^q, \underline{d}_{k-t}^q\}_{t=0}^k$ are the *augmented* input-output data set.

The state propagation step predicts the framers for the states at time step k using framers from the previous time step $k-1$. The approach is based on nonlinear bounding in Corollary 1 for the nonlinear function f^q in (6), where we take the tighter estimates between the over-approximation of propagation of f^q via decomposition functions and the ones obtained from the abstracted dynamics with $(A_k^{q,f}, \overline{e}_k^{q,f}, \underline{e}_k^{q,f})$ as the solution to the problem in (2a) for f^q.

In the attack policy learning step, the unknown function $d_k^q = \mu^q(\cdot)$ is learned/over-approximated, i.e., $\{\overline{\mu}_k^q(.), \underline{\mu}_k^q(.)\}$ is computed, by applying the nonparametric learning approach in [15, Theorem 1] using the augmented data set constructed from the estimated propagated framers, $\overline{x}_k^{q,p}, \underline{x}_k^{q,p}$, from the initial to the current time step. Note that the learning approach in [15, Theorem 1] guarantees that the true attack policy is contained in the approximated model, i.e., $\underline{\mu}_k^q(.) \leq d_k^q = \mu^q(.) \leq \overline{\mu}_k^q(.)$. Next, in the unknown input estimation step, the learned attack policy is used to find abstracted dynamics with $(A_k^{q,\mu}, \overline{e}_k^{q,\mu}, \underline{e}_k^{q,\mu})$ as the solution to the abstraction problem in (2a) for $\{\overline{\mu}_k^q(.), \underline{\mu}_k^q(.)\}$, where we again utilize the nonlinear bounding result in Corollary 1 to obtain the unknown input framers in (10).

Then, the measurement update step improves the propagated framers $[\underline{z}_k^{q,p}, \overline{z}_k^{q,p}]$ by "intersecting" them with the approximated inverse of the function g^q from the measurement/output equation in (6). To perform this *inversion*, we first find its affine abstraction using the abstraction problem in (2a) for g^q to obtain $(\mathbb{A}_k^{q,g} \triangleq \begin{bmatrix} A_k^{q,g} & W_k^{q,g} \end{bmatrix}, \overline{e}_k^{q,g}, \underline{e}_k^{q,g})$ and then we obtain all solutions $[\underline{z}_k^{q,u}, \overline{z}_k^{q,u}]$ of the resulting linear/affine equation using the Moore-Penrose pseudoinverse. Finally, since the abstraction problem in (2a) is dependent on its domain, the updated framers can be iteratively improved with shrinking domains and thus, we iteratively compute the *sequences of updated framers* $\{\overline{z}_{i,k}^{q,u}, \underline{z}_{i,k}^{q,u}\}_{i=1}^\infty$ as follows:

$$\begin{bmatrix} \overline{z}_{0,k}^{q,u} & \underline{z}_{0,k}^{q,u} \end{bmatrix} = \begin{bmatrix} \overline{z}_k^{q,p} & \underline{z}_k^{q,p} \end{bmatrix}, \quad \forall i \in \{1 \ldots \infty\}: \tag{12}$$

$$\begin{bmatrix} \overline{z}_{i,k}^{q,u} \\ \underline{z}_{i,k}^{q,u} \end{bmatrix} = \begin{bmatrix} \min(A_{i,k}^{q,g\dagger+} \overline{\alpha}_{i,k}^q - A_{i,k}^{q,g\dagger-} \underline{\alpha}_{i,k}^q + \omega_{i,k}^q, \overline{z}_{i-1,k}^{q,u}) \\ \max(A_{i,k}^{q,g\dagger+} \underline{\alpha}_{i,k}^q - A_{i,k}^{q,g\dagger-} \overline{\alpha}_{i,k}^q - \omega_{i,k}^q, \underline{z}_{i-1,k}^{q,u}) \end{bmatrix}, \tag{13}$$

where

$$\begin{bmatrix} \overline{\alpha}_{i,k}^q \\ \underline{\alpha}_{i,k}^q \end{bmatrix} = \begin{bmatrix} \min(\overline{t}_{i,k}^q, A_{i,k}^{q,g+} \overline{z}_{i-1,k}^{q,u} - A_{i,k}^{q,g-} \underline{z}_{i-1,k}^{q,u}) \\ \max(\underline{t}_{i,k}^q, A_{i,k}^{q,g+} \underline{z}_{i-1,k}^{q,u} - A_{i,k}^{q,g-} \overline{z}_{i-1,k}^{q,u}) \end{bmatrix}, \tag{14}$$

$$\begin{bmatrix} \overline{t}_{i,k}^q \\ \underline{t}_{i,k}^q \end{bmatrix} = \begin{bmatrix} y_k \\ y_k \end{bmatrix} + \begin{bmatrix} W_{i,k}^{q,g-} & -W_{i,k}^{q,g+} \\ -W_{i,k}^{q,g+} & W_{i,k}^{q,g-} \end{bmatrix} \begin{bmatrix} \overline{v} \\ \underline{v} \end{bmatrix} - \begin{bmatrix} \underline{e}_{i,k}^{q,g} \\ \overline{e}_{i,k}^{q,g} \end{bmatrix}, \tag{15}$$

$\omega_{i,k}^q = \kappa \text{rowsupp}(I - A_{i,k}^{q,g\dagger} A_{i,k}^{q,g})$, $\forall i \in \{1 \ldots \infty\}$ and κ is a very large positive real number (infinity).

2) Mode Estimation Observer: To estimate the set of compatible modes, we consider a membership-based elimination approach that checks if residual signals are within some compatible intervals. We first define the mode-matched residual signal r_k^q as the difference between the measured output y_k and the predicted output based on the predicted framers $[\underline{z}_k^{q,p}, \overline{z}_k^{q,p}]$ in (6) as follows.

Definition 5 (Residuals)**.** *For each mode q at time step k, the residual signal r_k^q is defined as:*

$$r_k^q \triangleq y_k - \frac{1}{2}(\overline{g}_k^q + \underline{g}_k^q), \tag{16}$$

where $\overline{g}_k^q, \underline{g}_k^q$ are values of bounding signals $\overline{g}^q, \underline{g}^q$ at time k, computed based on (4) *applied to the mapping $g^q(\cdot)$ in* (6) *with the predicted framer $[\underline{z}_k^{q,p}, \overline{z}_k^{q,p}]$ and $[\underline{v}, \overline{v}]$.*

Then, we eliminate a specific mode q, if its corresponding residual signal r_k^q satisfy a specific criterion, as follows:

Proposition 4 (Mode Elimination Criterion)**.** *Mode q is not a true mode if*

$$r_k^q \notin \mathcal{R}_k^q \triangleq \frac{1}{2}[-(\overline{g}_k^q - \underline{g}_k^q), \overline{g}_k^q - \underline{g}_k^q]. \tag{17}$$

Proof. If q is the true mode, then $y_k = g^q(x_k, d_k, v_k)$ by (6). Consequently, $y_k \in [\underline{g}_k, \overline{g}_k]$ which is equivalent to $r^q \in \mathcal{R}_k^q$, given the definition of r_k^q in (16), and with $\underline{g}_k, \overline{g}_k$ obtained from (4) in Corollary 1. ∎

By Proposition 4, if the residual signal of a particular mode q is not within the given interval in (17) conditioned on this mode being true, then q can be ruled out as incompatible.

3) Global Fusion Observer: Finally, combining the results above, our proposed global fusion observer will provide mode, attack and state set-valued estimates, as well as attack policy abstractions, at each time step k as:

$$\begin{gathered} \hat{\mathcal{Q}}_k = \{q \in \hat{\mathcal{Q}}_{k-1} \,|\, r_k^q \in \mathcal{R}_k^q\}, \\ \hat{\mathcal{X}}_k = \cup_{q \in \hat{\mathcal{Q}}_k} \mathcal{X}_k^q, \hat{\mathcal{D}}_k = \cup_{q \in \hat{\mathcal{Q}}_k} \mathcal{D}_k^q, \\ \overline{\mu}_k(\cdot) = \max_{q \in \hat{\mathcal{Q}}} \overline{\mu}_k^q(\cdot), \ \underline{\mu}_k(\cdot) = \min_{q \in \hat{\mathcal{Q}}} \underline{\mu}_k^q(\cdot). \end{gathered}$$

Algorithm 1 Simultaneous Mode, State and Attack Policy (SMSP) Estimation

$\hat{\mathcal{Q}}_0 = \mathcal{Q}$;
for $k = 1$ to N **do**
  for $q \in \hat{\mathcal{Q}}_{k-1}$ **do**
  ▷ Mode-Matched State and Attack Policy Set-Valued Estimates
    Compute $\overline{x}_k^q, \underline{x}_k^q, \overline{d}_k^q, \underline{d}_k^q$ through (8a)–(11b);
  ▷ Mode Observer via Elimination
    $\hat{\mathcal{Q}}_k = \hat{\mathcal{Q}}_{k-1}$;
    Compute r_k^q via Definition 5;
    if (17) holds **then** $\hat{\mathcal{Q}}_k = \hat{\mathcal{Q}}_k \backslash \{q\}$;
    end if
  end for
  ▷ State and Input Estimates
  $\hat{\mathcal{X}}_k = \cup_{q \in \hat{\mathcal{Q}}_k} \hat{\mathcal{X}}_k^q$; $\hat{\mathcal{D}}_k = \cup_{q \in \hat{\mathcal{Q}}_k} \hat{\mathcal{D}}_k^q$;
  ▷ Attack Policy Abstraction
  $\overline{\mu}_k(\cdot) = \max_{q \in \hat{\mathcal{Q}}} \overline{\mu}_k^q(\cdot)$; $\underline{\mu}_k(\cdot) = \min_{q \in \hat{\mathcal{Q}}} \underline{\mu}_k^q(\cdot)$;
end for

The simultaneous mode, state and attack policy (SMSP) estimation approach is summarized in Algorithm 1.

B. Properties of Mode-Matched Observers

In this section, following a similar approach to our previous work [12], we show that each of the mode-matched observers is correct (cf. Definition 3) and stable (cf. Definition 4) under some sufficient conditions. Moreover, the sequence of mode-matched interval widths is convergent to some computable steady state values.

Lemma 1 (Correctness)**.** *Consider System* (6) *and suppose Assumptions 1–3 hold. Then, for all mode $q \in \mathcal{Q}$, the dynamical system in* (8a)–(11b) *constructs a correct mode-matched interval observer for System* (6)*, conditioned on the mode being the true mode, i.e., $q = q^*$. In other words, $\forall k \in \mathbb{K} \triangleq \mathbb{N} \cup \{0\}, \underline{z}_k^q \leq z_k^q \leq \overline{z}_k^q$, where $z_k^q \triangleq [x^\top \ d^{q\top}]^\top$ and $[\underline{z}_k^q, \overline{z}_k^q]$ are the augmented vectors of state and unknown inputs in the dynamical systems in* (6) *and the updated framer from* (11a) *at time $k \in \mathbb{K}$, respectively.*

Proof. Using induction, the proof is similar to the proof of [12, Theorem 1]. ■

Next, we address the stability of each mode-matched observer. Note that similar to [12], our goal is to obtain sufficient stability conditions that can be checked *a priori* instead of for each time step k. On the other hand, for the implementation of the update step, we iteratively find new mode-matched *local* parallel abstraction slopes $A_{i,k}^{q,g}$ by iteratively solving the LP (2a) for g^q on the intervals obtained in the previous iteration, $\mathcal{B}_{i,k}^{q,u} = [\underline{z}_{i-1,k}^{q,u}, \overline{z}_{i-1,k}^{q,u}]$, to find *local* framers $\overline{z}_{i,k}^{q,u}, \underline{z}_{i,k}^{q,u}$ (cf. (12)–(14)), with additional constraints given in (2b) in the optimization problems, which guarantees that the iteratively updated *local* intervals obtained using the local abstraction slopes are inside the global interval, i.e.,

$$\underline{z}_k^{q,u} \leq \underline{z}_{0,k}^{q,u} \leq \cdots \leq \underline{z}_{i,k}^{q,u} \leq \cdots \leq \lim_{i \to \infty} \underline{z}_{i,k}^{q,u} \triangleq \underline{z}_k^q,$$
$$\overline{z}_k^q \triangleq \lim_{i \to \infty} \overline{z}_{i,k}^{q,u} \leq \cdots \leq \overline{z}_{i,k}^{q,u} \leq \cdots \leq \overline{z}_{0,k}^{q,u} \leq \overline{z}_k^{q,u}.$$

With that in mind, we next show through the following proposition that the sequence of the widths of the interval-valued estimates are upper bounded by a difference equation, i.e., a discrete-time dynamical system, for each mode.

Proposition 5 (Interval Widths Upper System)**.** *Consider System* (6) *along with the observer in* (8a)–(11b) *and suppose that all the assumptions in Lemma 1 hold and the decomposition function f_d is constructed using* (5)*. Define the mode-matched width of the interval-valued estimate $[\underline{z}_k^q, \overline{z}_k^q]$, at time k, as $\Delta_k^{z^q} \triangleq \overline{z}_k^q - \underline{z}_k^q$. Then, for each mode $q \in \mathcal{Q}$, the following holds: $\forall (D_1^q, D_2^q, D_3^q) \in \mathbb{D}_{n+p} \times \mathbb{D}_l \times \mathbb{D}_n$,*

$$\Delta_k^{z^q} \leq \mathcal{A}_q^g(D_1^q, D_2^q)\mathcal{A}_q^{f,h}(D_3^q)\Delta_{k-1}^{z^q} + \Delta_q^g(D_1^q, D_2^q) + \mathcal{A}_q^g(D_1^q, D_2^q)\Delta_q^{f,h}(D_3^q) + 2\kappa D_1^q \mathbf{r}^q, \tag{18}$$

where

$$\mathcal{A}_q^g(D_1^q, D_2^q) \triangleq D_1^q |A_q^{g\dagger}| D_2^q |A_q^g| + (I - D_1^q),$$
$$\mathcal{A}_q^{f,\mu}(D_3^q) \triangleq \left[(|A_q^f| + 2(I - D_3^q)C_z^{f^q})^\top \ [|A_q^\mu| \ 0]^\top \right]^\top,$$
$$\Delta_q^g(D_1^q, D_2^q) \triangleq D_1^q |A_q^{g\dagger}| D_2^q (|W_q^g| \Delta v + \Delta_e^{g^q}),$$
$$\Delta_q^{f,h}(D_3^q) \triangleq (|W_q^f| + 2(I - D_3^q)C_w^{f^q})\Delta w + \Delta_e^{f^q},$$

while $\mathbf{r}^q \triangleq \text{rowsupp}(I - A_q^{g\dagger} A_q^g)$, $C_q^f \triangleq \begin{bmatrix} C_z^{f^q} & C_u^{f^q} & C_w^{f^q} \end{bmatrix}$ *from* (5)*, κ is a very large positive real number (infinity) and* $\Delta_e^{g^q} \triangleq \overline{e}^{f^q} - \underline{e}^{f^q}, \Delta_e^{f^q} \triangleq \overline{e}^{g^q} - \underline{e}^{g^q}, \Delta v \triangleq \overline{v} - \underline{v}, \Delta w \triangleq \overline{w} - \underline{w}$, $\{A_q^s \triangleq \mathbb{A}_{(1:n+p)}^s\}_{s \in \{f^q, g^q\}}, A_q^\mu \triangleq \mathbb{A}^{\mu^q}, W_q^f \triangleq \mathbb{A}_{(n+p+1:n+p+n_w)}^s, W_q^g \triangleq \mathbb{A}_{(n+p+1:n+p+n_v)}^g$*, with* $\mathbb{A}^s$ *and* $\mathbb{A}^{\mu^q}$ *obtained using Proposition 2.*

Proof. The proof is similar to the proof of [12, Theorem 2], with some minor modifications, by replacing the unknown mapping h with the unknown policy μ and making all variables mode-dependent. ■

Now, armed with the results in Proposition 5, we provide sufficient conditions for the stability of each of the mode-matched observers in the sense of Definition 4, in a similar manner to [12, Theorem 2], through the following lemma.

Lemma 2 (Stability)**.** *Consider the hidden mode switched system* (6) *along with the mode-matched observer in* (8a)–(11b)*. Suppose that all the assumptions in Proposition 5 hold. Then, for each mode $q \in \mathcal{Q}$, the mode-matched observer in* (8a)–(11b) *is stable (cf. Definition 4), if there exist $D_1^q \in \mathbb{D}_{n+p}, D_2^q \in \mathbb{D}_l, D_3^q \in \mathbb{D}_n$ that satisfy $D_{1,i,i}^q = 0$ if $\mathbf{r}^q(i) = 1$, i.e., if there exist $(D_1^q, D_2^q, D_3^q) \in \mathbb{D}^* \triangleq \{(D_1, D_2, D_3) \in \mathbb{D}_{n+p} \times \mathbb{D}_l \times \mathbb{D}_n | D_{1,ii}\mathbf{r}(i) = 0\}$ such that*

$$\mathcal{L}^*(D_1^q, D_2^q, D_3^q) \triangleq \|\mathcal{A}_q^g(D_1^q, D_2^q)\mathcal{A}_q^{f,\mu}(D_3^q)\| \leq 1, \tag{19}$$

with $\mathcal{A}_q^g(D_1^q, D_2^q)$ and $\mathcal{A}_q^{f,\mu}(D_3^q)$ defined in Proposition 5.

Proof. Our goal is to show that our specific choices for D_1^q, D_2^q, D_3^q make the right hand side of (18) finite in finite time. To do this, since κ can be infinitely large, we choose $D_1^q \in \mathbb{D}_{n+p}$ such that $D_1^q \mathbf{r}^q = 0$, i.e., $D_{1,i,i}^q = 0$ if $\mathbf{r}^q(i) = 1, i = 1, \ldots, n+p$. Then, by the *comparison lemma* [21, Lemma 3.4], it suffices for uniform boundedness of $\{\Delta_k^{z^q}\}_{k=0}^\infty$ that the following system:

$$\Delta_k^{z^q} = \mathcal{A}_q^g(D_1^q, D_2^q)\mathcal{A}_q^{f,\mu}(D_3^q)\Delta_{k-1}^{z^q} + \tilde{\Delta}_q(D_1^q, D_2^q) \tag{20}$$

be stable, where $\tilde{\Delta}_q(D_1^q, D_2^q) \triangleq \Delta_q^g(D_1^q, D_2^q) + \mathcal{A}_q^g(D_1^q, D_2^q)\Delta_q^{f,\mu}(D_3^q)$ is a bounded disturbance by construction. This implies that the system (20) is stable (in the sense of uniform stability of the interval sequnces) if and only if the matrix $\mathcal{A}_q(D_1^q, D_2^q, D_3^q) \triangleq \mathcal{A}_q^g(D_1^q, D_2^q)\mathcal{A}_q^{f,\mu}(D_3^q)$ is (non-strictly) stable for at least one choice of (D_1^q, D_2^q, D_3^q), and equivalently, (19) should hold. ∎

Finally, the mode-matched interval widths are upper bounded and convergent to steady-state values, as follows.

Proposition 6 (Convergence of Upper Bounds of the Interval Widths)**.** *Consider System* (6) *and the observer* (8a)–(11b) *and suppose all assumptions in Lemma 2 hold. Then, for each mode $q \in \mathcal{Q}$, the sequence of $\{\Delta_k^{z^q} \triangleq \overline{z}_k^q - \underline{z}_k^q\}_{k=0}^{\infty}$ is uniformly upper bounded by a convergent sequence:*

$$\Delta_k^{z^q} \leq \overline{\mathcal{A}}_q^k \Delta_0^{z^q} + \textstyle\sum_{j=0}^{k-1} \overline{\mathcal{A}}_q^j \overline{\Delta}_q \xrightarrow{k\to\infty} e^{\overline{\mathcal{A}}_q} \overline{\Delta}_q,$$

where $\overline{\mathcal{A}}_q = \mathcal{A}_q(D_1^{q\star}, D_2^{q\star}, D_3^{q\star}) \triangleq \mathcal{A} - q^g(D_1^{q\star}, D_2^{q\star})\mathcal{A}_q^{f,\mu}(D_3^{q\star}), \overline{\Delta}_q = \Delta_q^g(D_1^{q\star}, D_2^{q\star}) + \mathcal{A}_q^g(D_1^{q\star}, D_2^{q\star})\Delta_q^{f,\mu}(D_3^{q\star})$, and $(D_1^{q\star}, D_2^{q\star}, D_3^{q\star})$ is a solution of the following problem:

$$\min_{D_1,D_2,D_3} \| e^{\mathcal{A}_q(D_1,D_2,D_3)}(\Delta_q^g(D_1, D_2) + \mathcal{A}_q^g(D_1, D_2)\Delta_q^{f,\mu}(D_3))\|$$
$$\text{s.t.} (D_1, D_2, D_3) \in \{(D_1, D_2, D_3) \in \mathbb{D}^* \,|\, \mathcal{L}_q^*(D_1, D_2, D_3) < 1\}.$$

Consequently, the interval widths $\{\|\Delta_k^{z^q}\|\}_{k=1}^{\infty}$ are uniformly upper bounded by a convergent sequence, i.e., $\|\Delta_k^{z^q}\| \leq \delta_k^{z^q} \triangleq \|\overline{\mathcal{A}}_q^k \Delta_0^{z^q} + \sum_{j=0}^{k-1} \overline{\mathcal{A}}_q^j \overline{\Delta}_q\| \xrightarrow{k\to\infty} \|e^{\overline{\mathcal{A}}_q}\overline{\Delta}_q\|$.

Proof. The proof is straightforward by applying [22, Lemma 1], computing (18) iteratively, using triangle inequality and the fact that by Theorem 2, $\mathcal{A}_q(D_1^{q\star}, D_2^{q\star}, D_3^{q\star})$ is a stable matrix and $(D_1^{q\star}, D_2^{q\star}, D_3^{q\star})$ is a solution of (19). ∎

V. Mode-Detectability

In addition to the nice properties regarding the correctness, stability and convergence of the mode-matched interval estimates of states and inputs, as discussed in the previous section, we now provide some sufficient conditions for the system dynamics and attack policies, which guarantee that regardless of the observations, after some large enough time steps, *all* the false (i.e., not true) modes can be eliminated, when applying Algorithm 1. To do so, first, we define the concept of mode-detectability and state some assumptions for deriving our sufficient conditions for mode-detectability.

Definition 6 (Mode-Detectability)**.** *System* (6) *is called mode-detectable under Algorithm 1, if there exists a natural number $K \in \mathbb{N}$, such that for all time steps $k \geq K$, all false modes are eliminated.*

Assumption 4 (Bounded Jacobians)**.** *For all $q \in \mathcal{Q}$, the vector fields $f^q(x, \mu^q(x), w)$ and $\mu^q(x)$ satisfy the following bounds on their Jacobians: $\forall j \in \{x, d\}, \forall (x, w) \in \mathcal{X} \times \mathcal{W}$, $J_j^{f^q}(x, \mu^q(x), w) \in [\underline{J}_j^{f^q}, \overline{J}_j^{f^q}]$ and $J^{\mu^q}(x) \in [\underline{J}^{\mu^q}, \overline{J}^{\mu^q}]$, with known $\overline{J}^{f^q}, \underline{J}^{f^q}, \overline{J}^{\mu^q}, \overline{J}^{\mu^q}$ a priori, where $J_x^{f^q}, J_d^{f^q}$ are Jacobians of f^q, with respect to its first argument, x, and second argument d, respectively.*

Assumption 5 (Destabilizing Attack Policy)**.** *The matrix $J_q^{\mathrm{m}} \triangleq \frac{1}{2}(\underline{J}_x^{f^q} + \overline{J}_x^{f^q} + \underline{J}_d^{f^q,\mu^q} + \overline{J}_d^{f^q,\mu^q})$ is strictly Schur unstable, where $\underline{J}_d^{f^q,\mu^q} \triangleq \underline{J}_d^{f^q+}\underline{J}^{\mu^q+} - \overline{J}_d^{f^q+}\underline{J}^{\mu^q-} - \underline{J}_d^{f^q-}\overline{J}^{\mu^q+} + \overline{J}_d^{f^q-}\overline{J}^{\mu^q-}$ and $\overline{J}_d^{f^q,\mu^q} \triangleq \overline{J}_d^{f^q+}\overline{J}^{\mu^q+} - \underline{J}_d^{f^q+}\overline{J}^{\mu^q-} - \overline{J}_d^{f^q-}\underline{J}^{\mu^q+} + \underline{J}_d^{f^q-}\underline{J}^{\mu^q-}$, i.e., J_q^{m} has at least one eigenvalue whose absolute value is strictly greater than 1.*

Corollary 2. *Assumption 5 implies that the attack policy $\mu(\cdot) \triangleq \mu^{q^*}(\cdot)$ destabilizes the system in* (6)*, and hence, x_k, the true state trajectory of* (6) *becomes unbounded.*

Proof. Defining $f \triangleq f^{q^*}$, $\mu(x) \triangleq \mu^{q^*}(x)$ and $\tilde{f}(x, w) \triangleq f(x, \mu(x), w)$, as well as using *chain rule*, we have $J_x^{\tilde{f}} = J_x^f(x, \mu(x), w) + J_d^f(x, \mu(x), w)J^{\mu}(x)$. Combining this and Assumption 5, as well as applying Proposition 1, returns $J_x^{\tilde{f}} \in [\underline{J}_x^f + \underline{J}_d^{f,\mu}, \overline{J}_x^f + \overline{J}_d^{f,\mu}]$. Now, note that since $J_{q^*}^{\mathrm{m}}$ is strictly Schur unstable by Assumption 5, then the interval matrix $J_x^{\tilde{f}}$ is strictly Schur unstable by [23, Theorem 1a], and hence, the linearized form of the system in (6) is strictly Schur unstable. Consequently, the nonlinear system in (6) is unstable by the Chetaev instability theorem [24], i.e., the attack policy μ is a destabilizing policy. ∎

Now, we are ready to state our main result on mode-detectability, through the following theorem.

Theorem 1 (Sufficient Conditions for Mode-Detectability)**.** *Suppose Assumptions 4 and 5 and all the assumptions in Lemma 2 hold for all $q \in \mathcal{Q}$. Then, using Algorithm 1, System* (6) *is mode-detectable in the sense of Definition 6.*

Proof. We need to show that there exists $K \in \mathbb{N}$, such that (17) holds for all $k \geq K, \forall q \neq q^* \in \mathcal{Q}$, where q^* is the true mode. Given the definition of the residual signal in (16) and since q^* is unknown, a sufficient condition for (17) to hold is that $\forall q_1 \neq q_2 \in \mathcal{Q}, \exists K \in \mathbb{N}, \forall k \geq K, g^{q_2}(\xi_k^{q_2}) \notin [\underline{g}_k^{q_1}, \overline{g}_k^{q_1}]$, where $\forall q \in \mathcal{Q}, \xi_k^q \triangleq [x_k^\top \; d_k^{q\top} \; v_k^{q\top}]$. Equivalently, there should exist a dimension $i \in \mathbb{N}_l$, such that

$$g_i^{q_2}(\xi_k^{q_2}) < \underline{g}_{i,k}^{q_1} \text{ or } g_i^{q_2}(\xi_k^{q_2}) > \overline{g}_{i,k}^{q_1}. \tag{21}$$

Since $q_1 \neq q_2$ can be any two arbitrary modes, then without loss of generality, we only consider the former inequality in (21) that holds, if

$$\overline{g}_{i,k}^{q_2} < \underline{g}_{i,k}^{q_1}, \tag{22}$$

since $\overline{g}_{i,k}^{q_2}$ is an over-approximation for $g_i^{q_2}(\xi_k^{q_2})$. Further, by defining $\overline{\Delta}_{g_i}^q \triangleq \overline{g}_{i,k}^q - g_i^q(\xi_k^q) \geq 0$ and $\underline{\Delta}_{g_i}^q \triangleq g_i^q(\xi_k^q) - \underline{g}_{i,k}^q \geq 0$, (22) is equivalent to $g_i^{q_2}(\xi_k^{q_2}) + \overline{\Delta}_{g_i}^{q_2} < g_i^{q_1}(\xi_k^{q_1}) - \underline{\Delta}_{g_i}^{q_1}$, that can be rewritten as:

$$\overline{\Delta}_{g_i}^{q_2} + \underline{\Delta}_{g_i}^{q_1} \leq g_i^{q_1}(\xi_k^{q_1}) - g_i^{q_2}(\xi_k^{q_2}). \tag{23}$$

Note that the left hand side of (23) can be verified to be bounded as follows: $0 \leq \overline{\Delta}_{g_i}^{q_2} + \underline{\Delta}_{g_i}^{q_1} \leq \Delta_{g_i}^{q_2} + \Delta_{g_i}^{q_1}$, where $\forall q \in \mathcal{Q}, \Delta_{g_i}^q \triangleq \overline{g}_{i,k}^q - \underline{g}_{i,k}^q$ is bounded by the the Lipschitz-like property of the decomposition functions (cf. [12, Lemma 2]), and the stability of each of the mode-matched observers (cf. Lemma 2). Now that the left hand side of (23) is proven to be bounded, if we show that the right hand side grows unboundedly, then the inequality in (23) must always hold after some sufficiently large time step $K \in \mathbb{N}$. To do so, we consider some $x_0 \in \mathcal{X}_0$ and apply the mean value theorem on both $g_i^{q_1}$ and $g_i^{q_2}$ to obtain

$$g_i^{q_1}(\xi_k^{q_1}) - g_i^{q_2}(\xi_k^{q_2}) = (J_{g,i}^{q_2}(\tilde{\xi}_k^2) - J_{g,i}^{q_1}(\tilde{\xi}_k^1))(x_k - x_0), \quad (24)$$

where $\tilde{\xi}_k^2, \tilde{\xi}_k^1 \in \mathcal{X} \times \mathbb{R}^p \times \mathcal{V}$ and for all $q \in \mathcal{Q}$, $J_g^q(\xi)$ and $J_{g,i}^q(\xi)$ denote the Jacobian matrix of g^q and its i'th row, evaluated at ξ. Finally, the right hand side of (24) eventually becomes unbounded, since the Jacobian matrix of J^q is bounded for all $q \in \mathcal{Q}$ by Lipschitz continuity of g^q (cf. Assumption 1), and x_k, i.e., the true state trajectory becomes unbounded by Assumption 5 and Corollary 2. Hence, the left hand side of the above equality also becomes eventually unbounded, which returns the desired result. ■

VI. Simulation Example

In this section, we illustrate the effectiveness our proposed observer using a power network with multiple control areas. Specifically, we consider a 3-area system as shown in Figure 1 where each control area consists of generator and load buses. In addition, there are transmission lines between areas. The nonlinear model of bus i is adopted from [25]:

$$\dot{\theta}_i(t) = f_i(t) + w_{1,i}(t),$$

$$\dot{f}_i(t) = -\frac{D_i f_i(t) + \sum_{l \in S_i} P_{il}(t) - (P_{M_i}(t) + d_i(t)) + P_{L_i}(t)}{m_i} + w_{2,i}(t),$$

with the output model: $y_{i,k} = [\theta_{i,k}, f_{i,k}]^\top + [0,1]^\top d_i + v_{i,k}$, where θ_i is phase angle, f_i is the angular frequency, $P_{M_i}(t)$ is the mechanical power (the control input), $P_{L_i}(t)$ is a known power demand, and S_i is the set of neighboring buses of i. In our simulations, both $P_{M_i}(t)$ and $P_{L_i}(t)$ are set to be identically zero and the process noise $w_i(t)$ and measurement noise $v_i(t)$ are both bounded by $\left[\begin{bmatrix}-0.1\\-0.1\end{bmatrix}, \begin{bmatrix}0.1\\0.1\end{bmatrix}\right]$. When the circuit breakers are not engaged (or attacked), the power flow P_{il} between areas i and l is as follows:

$$P_{il}(t) = -P_{li}(t) = t_{il} \sin(\theta_i(t) - \theta_j(t)).$$

A malicious agent is assumed to have access to circuit breakers that control the tie-lines, and is thus able to sever the connection between control areas. Two types of attack are considered based on the topology of the tie-line interconnection graph: (1) a node/vertex/bus attack (disconnection of a control area from all others); or (2) a link/edge/line attack (disabling of a specific tie-line between two control areas), i.e., the power flow across the tie lines is altered, if (1) there is an attack on control area i (node/bus attack): $P_{il}(t) = -P_{li}(t) = 0,\ \forall l \neq i$; or (2) if there exists an attack on circuit breaker (i,l) (link/line attack): $P_{il}(t) = -P_{li}(t) = 0$.

For the radial tie-line interconnection topology in Figure 1, the circuit breaker attacks result in $Q = 5$ possible modes of operation: all switches are safe ($q = 1$), only circuit breaker i is attacked ($q = i + 1,\ i = 1,2,3$) and two or more circuit breakers are attacked ($q = 5$). Further, we denote the value of the variables at sampling time t_k by adding subscript k, e.g., $f_i(t_k) = f_{i,k}$ and apply the Euler method to discretize the system: $\theta_{i,k+1} = \theta_{i,k} + \dot{\theta}_{i,k} dt,\ f_{i,k+1} = f_{i,k} + \dot{f}_{i,k} dt$, where the sampling time dt is $0.01s$ in our example. Moreover, we choose $d_i(t) = \theta_i(t)\sin(\theta_i(t))$ as to-be-learned unknown nonlinear attack policy and assume that we have 400 initial data points for each unknown attack policy d_i.

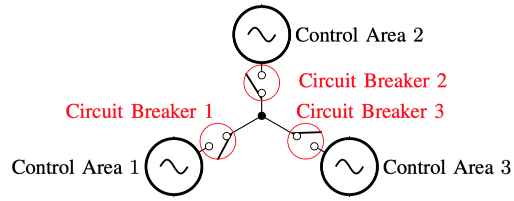

Fig. 1: Example of a three-area power station in a radial topology (corresponding to node/bus attack).

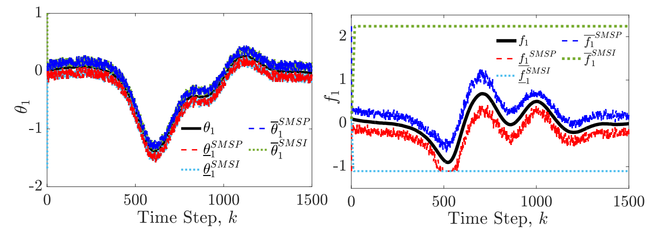

Fig. 2: The true values of the states: θ_1, f_1, and their upper and lower framers returned by the SMSP approach: $\overline{\theta}_1^{SMSP}, \underline{\theta}_1^{SMSP}, \overline{f}_1^{SMSP}, \underline{f}_1^{SMSP}$, as well as the SMSI approach: $\overline{\theta}_1^{SMSI}, \underline{\theta}_1^{SMSI}, \overline{f}_1^{SMSI}, \underline{f}_1^{SMSI}$.

Due to space limitations, we only show the results for the case when the true operation mode is assumed to be $q^* = 1$ and provide figures for selected states and attack signals in Figure 2 and 3. Moreover, we compare our results with our previously developed simultaneous mode, state and unknown input (SMSI) observer in [13], where no unknown policy/feedback law was assumed to govern the attack signals, and hence, no learning step were included in the proposed observer design. As can be observed from Figures 2 and 3, the SMSP observer (proposed in this paper) returns tighter interval estimates than SMSI for both states and attack signals, when compared to SMSI. It is also worth mentioning that all the state interval widths converge to steady state values by using SMSP with "learned" model for the attack policy, while the interval widths for some states do not converge when applying SMSI (not depicted for brevity), which highlights the effectiveness of the learning step.

Moreover, we compare the upper and lower learned model abstractions for $k = 0$ and $k = 1500$ in Figure 4, which showed tighter over-approximations with increasing number of data points. Further, as can be seen in Figure 5, all modes, except the true mode $q^* = 1$, are eliminated within 1500 time steps. Finally, the actual state and input estimation error sequence, as shown in Figure 6, is upper bounded by the interval widths and converges to steady-state values.

VII. Conclusion

This paper addresses the problem of designing interval observers for hidden mode switched nonlinear systems with bounded noise signals that are compromised by false data injection and switching attacks. An interval observer with three constituents was proposed: i) a bank of mode-matched observers, where each of them simultaneously outputs the corresponding mode-matched state, mode and unknown in-

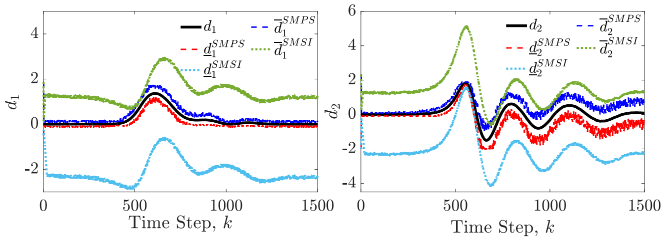

Fig. 3: The true values of the attacks: d_1, d_2, and their upper and lower framers returned by the SMSP approach: $\overline{d}_1^{SMSP}, \underline{d}_1^{SMSP}, \overline{d}_2^{SMSP}, \underline{d}_2^{SMSP}$, as well as the SMSI approach: $\overline{d}_1^{SMSI}, \underline{d}_1^{SMSI}, \overline{d}_2^{SMSI}, \underline{d}_2^{SMSI}$.

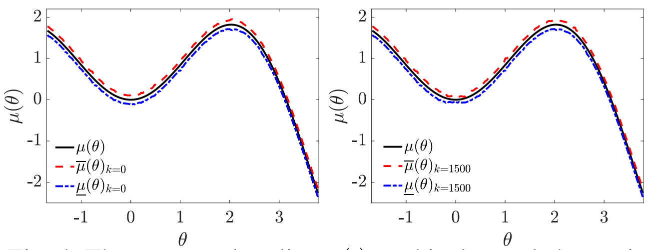

Fig. 4: The true attack policy $\mu(\cdot)$, and its learned abstraction model $\{\underline{\mu}(\cdot), \overline{\mu}(\cdot)\}$ at time steps $k = 0$ and $k = 1500$.

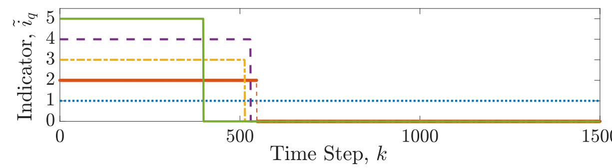

Fig. 5: Mode estimates with indicators $\tilde{i}_q = qi_q$, where $i_q = 0$ if mode q is eliminated and $i_q = 1$ otherwise.

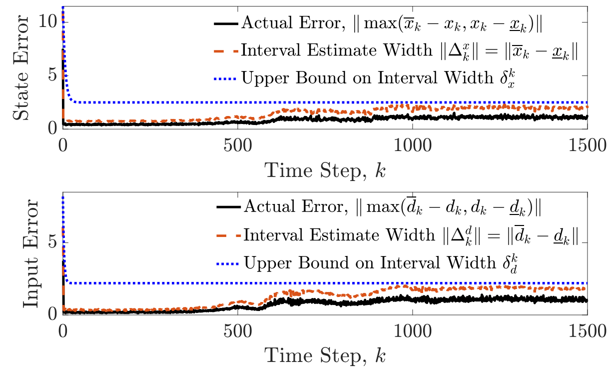

Fig. 6: State and input estimation error sequences.

put/attack estimates, as well as computes upper and lower abstractions/over-approximations for the attack policies, ii) a mode estimator that rules out the incorrect modes based on a residual-based set-membership criterion, and iii) a global fusion observer that returns the union of compatible state and attack estimates, as well as learned abstractions of the attack policy/state feedback law. Moreover, sufficient conditions for mode-detectability, i.e., for guaranteeing that all false modes will be eliminated after sufficiently large finite time steps, were provided. Finally, the effectiveness of our proposed approach was demonstrated using a 3-area power network.

References

[1] S.Z. Yong, M. Zhu, and E. Frazzoli. Switching and data injection attacks on stochastic cyber-physical systems: Modeling, resilient estimation, and attack mitigation. *ACM Transactions on Cyber-Physical Systems*, 2(2):9, 2018.

[2] Y. Bar-Shalom, T. Kirubarajan, and X.R. Li. *Estimation with Applications to Tracking and Navigation*. John Wiley & Sons, Inc., New York, NY, USA, 2002.

[3] F. Blanchini and M. Sznaier. A convex optimization approach to synthesizing bounded complexity ℓ^∞ filters. *IEEE Transactions on Automatic Control*, 57(1):216–221, 2012.

[4] M. Khajenejad and S.Z. Yong. Simultaneous state and unknown input set-valued observers for some classes of nonlinear dynamical systems. *arXiv preprint arXiv:2001.10125*, 2020.

[5] J. Giraldo, D. Urbina, A. Cardenas, J. Valente, M. Faisal, J. Ruths, N.O. Tippenhauer, H. Sandberg, and R. Candell. A survey of physics-based attack detection in cyber-physical systems. *ACM Computing Surveys (CSUR)*, 51(4):1–36, 2018.

[6] Y. Nakahira and Y. Mo. Attack-resilient $\mathcal{H}_2$, $\mathcal{H}_\infty$, and ℓ_1 state estimator. *IEEE Transactions on Automatic Control*, 63(12):4353–4360, 2018.

[7] 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 Transactions on Automatic Control*, 64(3):1162–1169, 2018.

[8] M.S. Chong, H. Sandberg, and J.P. Hespanha. A secure state estimation algorithm for nonlinear systems under sensor attacks. In *IEEE Conference on Decision and Control*, pages 5743–5748, 2020.

[9] M. Khajenejad and S.Z. Yong. Simultaneous mode, state and input set-valued observers for switched nonlinear systems. *arXiv preprint arXiv:2102.10793*, 2021.

[10] M. Milanese and C. Novara. Set membership identification of nonlinear systems. *Automatica*, 40:957–975, 2004.

[11] Z.B Zabinsky, R.L Smith, and B.P Kristinsdottir. Optimal estimation of univariate black-box Lipschitz functions with upper and lower error bounds. *Computers & Operations Res.*, 30(10):1539–1553, 2003.

[12] M. Khajenejad, Z. Jin, and S.Z. Yong. Interval observers for simultaneous state and model estimation of partially known nonlinear systems. In *American Control Conference*, pages 2848–2854. IEEE, 2021.

[13] M. Khajenejad and S. Z. Yong. Simultaneous mode, input and state set-valued observers with applications to resilient estimation against sparse attacks. In *IEEE Conference on Decision and Control (CDC)*, pages 1544–1550, 2019.

[14] M. Khajenejad and S.Z. Yong. Simultaneous input and state interval observers for nonlinear systems with full-rank direct feedthrough. In *IEEE Conference on Decision and Control*, pages 5443–5448, 2020.

[15] Z. Jin, M. Khajenejad, and S.Z. Yong. Data-driven model invalidation for unknown lipschitz continuous systems via abstraction. In *American Control Conference (ACC)*, pages 2975–2980. IEEE, 2020.

[16] D. Efimov, L. Fridman, T. Raissi, A. Zolghadri, and R. Seydou. Interval estimation for lpv systems applying high order sliding mode techniques. *Automatica*, 48(9):2365–2371, 2012.

[17] K.R. Singh, Q. Shen, and S.Z. Yong. Mesh-based affine abstraction of nonlinear systems with tighter bounds. In *IEEE Conference on Decision and Control (CDC)*, pages 3056–3061, 2018.

[18] L. Yang, O. Mickelin, and N. Ozay. On sufficient conditions for mixed monotonicity. *IEEE Transactions on Automatic Control*, 64(12):5080–5085, 2019.

[19] S. Coogan and M. Arcak. Efficient finite abstraction of mixed monotone systems. In *Hybrid Systems: Computation and Control*, pages 58–67. ACM, 2015.

[20] M. Khajenejad and S.Z. Yong. Tight remainder-form decomposition functions with applications to constrained reachability and interval observer design. *arXiv preprint arXiv:2103.08638*, 2021.

[21] H.K. Khalil. Nonlinear systems. *Upper Saddle River*, 2002.

[22] D. Efimov, T. Raïssi, S. Chebotarev, and A. Zolghadri. Interval state observer for nonlinear time varying systems. *Automatica*, 49(1):200–205, 2013.

[23] K. Wang, A.N. Michel, and D. Liu. Necessary and sufficient conditions for the hurwitz and schur stability of interval matrices. *IEEE Transactions on Automatic Control*, 39(6):1251–1255, 1994.

[24] N.G. Chetaev. Stability of motion: works on analytical mechanics. *AS USSR.–535.(in Russian)*, 1962.

[25] H. Kim, P. Guo, M. Zhu, and P. Liu. Attack-resilient estimation of switched nonlinear cyber-physical systems. In *American Control Conference (ACC)*, pages 4328–4333. IEEE, 2017.