Chapter 7
Resilient State Estimation and Attack Mitigation in Cyber-Physical Systems

Mohammad Khajenejad and Sze Zheng Yong

7.1 Introduction

Cyber-Physical Systems (CPS), e.g., power grids, autonomous vehicles, medical devices, etc., are systems in which computational and communication components are deeply intertwined and interacting with each other in several ways to control physical entities. While the cyber-physical coupling introduces new functions to control systems and improves their performance, these systems also become exposed to new cyber-vulnerabilities. Such *safety-critical* systems, if jeopardized or malfunctioning, can cause serious detriment to their operators and users, as well as the controlled physical components. A need for CPS security and for new designs of resilient estimation, attack mitigation and control has been accentuated by recent incidents of attacks on CPS, e.g., the Iranian nuclear plant, the Ukrainian power grid, and the Maroochy water service (Cárdenas et al. 2008; Farwell and Rohozinski 2011; Richards 2008; Slay and Miller 2007; Zetter 2016). Specifically, mode and false data injection attacks are among the most serious types of attacks on CPS, where malicious and/or strategic attackers compromise the true mode (i.e., discrete state) of the system and/or inject counterfeit data signals into the sensor measurements and actuator signals to cause damage, steal energy, etc. Hence, reliable estimates of modes, (continuous) states, and unknown inputs (attacks) are indispensable and useful for the sake of attack identification and mitigation and resilient control. Similar state and input estimation problems can be found across a wide range of disciplines, from input estimation in physiological systems (De Nicolao et al. 1997), to fault detection and diagnosis (Patton et al. 1989), to the estimation of mean areal precipitation (Kitanidis 1987).

M. Khajenejad · S. Z. Yong (✉)
Arizona State University, Tempe, AZ 85287, USA
e-mail: szyong@asu.edu

M. Khajenejad
e-mail: mkhajene@asu.edu

M. Abbaszadeh and A. Zemouche (eds.), *Security and Resilience in Cyber-Physical Systems*, https://doi.org/10.1007/978-3-030-97166-3_7

7.1.1 Literature Review

Characterization of undetectable attacks as well as attack detection and identification techniques have been extensively studied in the literature, which range from data-driven approaches (e.g., the use of data time-stamps in Zhu and Martínez (2013), Wasserstein metric in Li and Martínez (2020) or higher-order moments in Renganathan et al. 2021) to the works seeking closed-form solutions for selecting various types of detector thresholds (e.g., Murguia and Ruths 2016; Milošević et al. 2018) to anomaly detection methods using residuals (e.g., Mo and Sinopoli 2010; Weimer et al. 2012; Kwon et al. 2013) with empirically chosen thresholds to trade-off between false alarms and probability of anomaly/attack detection. On the other hand, attack mitigation can be preventive and/or reactive (Cómbita et al. 2015). Preventive attack mitigation identifies and removes system vulnerabilities to prevent exploitation (e.g., Dan and Sandberg 2010), while reactive attack mitigation, which is mainly studied using either game theory (e.g., Ma et al. 2013; Zhu and Martínez 2011; Zhu and Basar 2015) or adaptive learning and control architectures for mitigating sensor and actuator attacks (e.g., Jin et al. 2017; Yadegar et al. 2019; Jin and Haddad 2019, 2020), initiates countermeasures after detecting an attack.

The ability to reliably estimate the true system states despite attacks (i.e., resilient estimates) is also desirable in addition to attack detection or the resulting attack mitigation, because the availability of resilient state estimates would allow for continued operation with the same controllers as in the case without attacks or for pricing/prediction based on the real unbiased/compatible state information despite attacks. This problem has been addressed for both static systems (e.g., Liu et al. 2011; Kosut et al. 2011; Liang et al. 2017 and references therein) and dynamic systems (e.g., Mishra et al. 2015; Cárdenas et al. 2008; Mo and Sinopoli 2010; Pasqualetti et al. 2013; Fawzi et al. 2014; Pajic et al. 2014, 2015; Yong et al. 2016a; Dahleh and Diaz-Bobillo 1994; Shamma and Tu 1999; Blanchini and Sznaier 2012; Yong 2018; Yong et al. 2018).

In particular, resilient state estimators for *deterministic* linear dynamic systems under actuator and sensor signal attacks (e.g., via false data injection Cárdenas et al. 2008; Mo and Sinopoli 2010; Pasqualetti et al. 2013), have been proposed as a relaxed ℓ_0 optimization problem in Fawzi et al. (2014), and extensions in Pajic et al. (2014), Pajic et al. (2015) compute the worst-case bound on the state estimate errors in the presence of additive noise errors with known bounds, while Yong et al. (2016a) propose the resilient state estimators that are robust to bounded multiplicative and additive modeling and noise errors. On the other hand, our previous work Yong et al. (2015), Yong (2018) proposed to use a simultaneous input and state estimation (see, e.g., Yong 2018; Gillijns and De Moor 2007a, b; Yong et al. 2016b, 2017) approach for resilient state estimation , where we modeled the data injection attacks as unknown inputs of dynamical systems and derived stability and optimality properties for our estimators, as well as their relationship to strong detectability (Yong et al. 2016b).

In addition, a serious CPS security concern has emerged more recently from the attacks that alter the CPS network topology or exploit the switching vulnerability of CPS, e.g., attacks on the power system network topology (Weimer et al. 2012), or on the circuit breakers of a smart grid (Liu et al. 2013), on the meter/sensor data network topology (Kim and Tong 2013) or on the logic mode (e.g., failsafe mode) of a traffic infrastructure (Ghena et al. 2014). To address this concern, our previous works (Yong et al. 2021, 2018; Khajenejad and Yong 2019) proposed inference algorithms that estimate hidden modes, unknown inputs (attacks) and states simultaneously as a means to obtain resilient state estimation despite switching (mode/topology) attacks as well as attacks on actuator and sensor signals. This framework is inspired by the *multiple-model* approach (see e.g., Bar-Shalom et al. 2004; Mazor et al. 1998 and references therein) and can be viewed as a generalization of the robust control-inspired approach in Nakahira and Mo (2018) that considers resilient state estimation against sparse data injection attacks on only the sensors.

In the context of reactive attack mitigation, the work in Ma et al. (2013) utilized a Markov game analysis for attack-defense in power systems, while a leader–follower (Stackelberg) game formulation was developed in Zhu and Martínez (2011) to model the interdependency between the operator and adversaries and solved using a receding-horizon Stackelberg control law to maintain the closed-loop system stability and some performance specifications. Further, a cross-layer coupled design was presented in a hybrid game-theoretic framework in Zhu and Basar (2015), where the occurrence of unanticipated events was modeled by stochastic switching , and deterministic uncertainties were represented by disturbances with a known range, and a robust controller was then designed at the physical layer to take into account risks of failures due to the cyber-system.

In this chapter, assuming different models for uncertainties/noise signals, we propose resilient state estimation algorithms that output reliable estimates of the true system states despite false data injection attacks and switching attacks. Our resilient estimation algorithms address switching attacks as well as actuator and sensor attacks in the presence of stochastic and/or set-valued noise signals. Our approach is built upon a general purpose inference algorithm developed and applied in our previous works (Yong et al. 2021, 2018; Khajenejad and Yong 2019) for hidden-mode stochastic/bounded error switched linear systems with unknown inputs (attacks). We model switching and false data injection attacks on Cyber-Physical Systems (CPS) in the presence of stochastic/distribution-free noise signals as an instance of this system class. By doing so, we show that unbiased and set-valued state estimates (i.e., resilient state estimates) can be (asymptotically) recovered with the algorithms in Yong et al. (2021), Khajenejad and Yong (2019). Secondly, we characterize fundamental limitations to resilient estimation that is useful for preventative mitigation, such as the upper bound on the number of correctable/tolerable attacks, and consider the subject of attack detection. In addition, we provide sufficient conditions for designing unidentifiable attacks (from the attacker's perspective) and also sufficient conditions to obtain resilient state estimates even when the attacks are not identified (from the system operator/defender's perspective). Finally, we design an attack-mitigating and stabilizing dynamic $\mathscr{H}_\infty$-controller that contributes to the literature on non-game-theoretic reactive attack mitigation.

An earlier manuscript appeared in Yong et al. (2018), where we addressed the resilient state estimation problem under switching and false data injection attacks for *stochastic* hidden-mode CPS only, while in this chapter, we also consider the uncertainties that are set-valued and further present a novel *dynamic* $\mathscr{H}_\infty$-optimal controller design for attack mitigation. Further, we provide *necessary* conditions for the attack signal to be unidentifiable to add to the previously derived sufficient conditions in Yong et al. (2018).

Notation: $\mathbb{R}^n$ denotes the n-dimensional Euclidean space and $\mathbb{N}$ nonnegative integers. For a vector $v \in \mathbb{R}^n$ and a matrix $M \in \mathbb{R}^{p \times q}$, $\|v\|_2 \triangleq \sqrt{v^\top v}$, $\|v\|_\infty \triangleq \max\limits_{1 \le i \le n} | v_i |$ and $\|M\|_2$, and $\sigma_{\min}(M)$ denote their induced 2-norm and non-trivial least singular value, respectively.

7.2 Problem Formulation

7.2.1 Attack Modeling

Similar to Yong et al. (2018), two different classes of possibly time-varying attacks on Cyber-Physical Systems (CPS) are considered:

Data Injection Attacks: Attacks on actuator and sensor signals via manipulation or injection with "false" signals of unknown *magnitude* and *location* (i.e., subset of attacked actuators or sensors). In other words, signal attacks consist of both *signal magnitude attacks* and *signal location attacks*. *Examples:* Denial-of-service, deceptive attacks via data injection (Cárdenas et al. 2008; Pasqualetti et al. 2013).

Switching Attacks: Attacks on the switching mechanisms that change the system's *mode* of operation, or on the sensor data or interconnection network *topology*, which we will also refer to as *mode attacks*. *Examples:* Attack on circuit breakers (Liu et al. 2013), power network topology (Weimer et al. 2012), sensor data network (Kim and Tong 2013) and logic switch of a traffic infrastructure (Ghena et al. 2014).

7.2.1.1 Data Injection Attacks

For clarity, we assume for the moment that there is only one mode of operation, and that the linear system dynamics is not perturbed by any noise signals:

$$x_{k+1} = A_k x_k + B_k(u_k + d_k^a), \quad y_k = C_k x_k + D_k(u_k + d_k^a) + d_k^s,$$

where $x_k \in \mathbb{R}^n$ is the continuous state, $y_k \in \mathbb{R}^\ell$ is the sensor output, $u_k \in \mathbb{R}^m$ is the known input, $d_k^a \in \mathbb{R}^m$ and $d_k^s \in \mathbb{R}^\ell$ are attack signals that are injected into the actuators and sensors, respectively. The attack signals are sparse, i.e., if sensor $i \in \{1, \ldots, \ell\}$ is not attacked then necessarily $d_k^{s,(i)} = 0$ for all time steps k; otherwise

$d_k^{s,(i)}$ can take any value. Since we do not know which sensor is attacked, we refer to this uncertainty as the *signal location attack*, and the arbitrary values that $d_k^{s,(i)}$ can take as the *signal magnitude attack*. This holds similarly for attacks on actuators d_k^a.

If we have additional knowledge of which of the actuators and sensors are vulnerable to data injection attacks, we will use $\overline{G}_k$ and $\overline{H}_k$ to incorporate this information, resulting in the following system dynamics

$$x_{k+1} = A_k x_k + B_k u_k + \overline{G}_k d_k^a, \quad y_k = C_k x_k + D_k u_k + \overline{D}_k d_k^a + \overline{H}_k d_k^s.$$

If no such information is available, $\overline{G}_k = B_k$, $\overline{D}_k = D_k$, and $\overline{H}_k = I$. Further, in some cases, the actuator and sensor attack signals are coupled and cannot be separated. In order to take this into consideration, we represent the potentially coupled attack signals with d_k and introduce corresponding G_k and H_k matrices to obtain

$$x_{k+1} = A_k x_k + B_k u_k + G_k d_k, \quad y_k = C_k x_k + D_k u_k + H_k d_k.$$

The special case where the actuator and sensor attack signals are independent can be obtained with $d_k = \left[(d_k^a)^\top \; (d_k^s)^\top\right]^\top$, $G_k = \left[\overline{G}_k \; 0\right]$ and $H_k = \left[\overline{D}_k \; \overline{H}_k\right]$, which will be made more precise in Sect. 7.2.2.

7.2.1.2 Switching Attacks

A system may have multiple modes of operation, denoted by the set $\mathscr{Q}^m$ of cardinality $t_m \triangleq |\mathscr{Q}^m|$, due to the presence of switching mechanisms or different configurations/topologies of the sensor data or interconnection network, where each mode $q \in \mathscr{Q}^m$ has its corresponding set of system matrices, $\{A_k^q, B_k^q, C_k^q, D_k^q, G_k^q, H_k^q\}$. A *switching attack* or *mode attack* then refers to the change of the mode of operation q by an adversary without the knowledge of the system operator/defender.

7.2.1.3 Attacker Model Assumptions

The malicious *signal magnitude attack* may be a signal of any type (random or strategic) or model, and we assume that no 'useful' knowledge of the dynamics of d_k is available (uncorrelated with $\{d_\ell\}$ for all $k \neq \ell$, $\{w_\ell\}$ and $\{v_\ell\}$ for all ℓ).

7.2.2 System Description

Our role as a system operator/defender is to obtain resilient/reliable state estimates. Thus, we model the system in a way that facilitates this. In other words, we model the switching and false data injection attacks on a "noisy" dynamic system using a *hidden-mode switched linear discrete-time system with unknown inputs* (i.e., a dynamical system with multiple modes of operation where the system dynamics in each mode is linear and *uncertain*, and the mode and some inputs are not known/measured):

$$\begin{aligned}(x_{k+1}, q_k) &= (A_k^q x_k + B_k^q u_k^q + G_k^q d_k^q + w_k^q, q), && x_k \in \mathscr{C}_q,\\ (x_k, q)^+ &= (x_k, \delta^q(x_k)), && x_k \in \mathscr{D}_q,\\ y_k &= C_k^q x_k + D_k^q u_k^q + H_k^q d_k^q + v_k^q, &&\end{aligned} \tag{7.1}$$

where $x_k \in \mathbb{R}^n$ is the continuous system state and $q \in \mathscr{Q} = \{1, 2, \dots, \mathfrak{N}\}$ is the hidden discrete state or *mode*, which a malicious attacker can influence, while $\mathscr{C}_q$ and $\mathscr{D}_q$ are flow and jump sets, and $\delta^q(x_k)$ is the mode transition function. More details on the hybrid systems formalism can be found in Goebel et al. (2009). For each mode q, $u_k^q \in U_q \subset \mathbb{R}^m$ is the known input, $d_k^q \in \mathbb{R}^p$ the unknown input or *attack signal*[1] and $y_k \in \mathbb{R}^l$ the output, whereas the corresponding process noise $w_k^q \in \mathbb{R}^n$ and measurement noise $v_k^q \in \mathbb{R}^l$ satisfy one of the following sets of assumptions for the system uncertainties:

Assumption 7.1 (*Aleatoric Uncertainty*) The system is perturbed by random (unbounded) process and measurement noise signals with process noise w_k^q and measurement noise v_k^q that are mutually uncorrelated, zero-mean Gaussian white random signals with known covariance matrices, $Q_k^q = \mathbb{E}[w_k^q w_k^{q\top}] \succeq 0$ and $R_k^q = \mathbb{E}[v_k^q v_k^{q\top}] \succ 0$, respectively. Moreover, x_0 is independent of v_k^q and w_k^q for all k.

Assumption 7.2 (*Epistemic Uncertainty*) The system is perturbed by uncertain, bounded process and measurement noise signals, where the corresponding process noise w_k^q and measurement noise v_k^q are distribution-free uncertain bounded signals with known bounds, i.e., $\|w_k^q\| \leq \eta_w$ and $\|v_k^q\| \leq \eta_v$, respectively (thus, they are ℓ_∞ sequences), where η_w^q and η_v^q are known parameters. We also assume an estimate $\hat{x}_0$ of the initial state x_0 is available, where $\|\hat{x}_0 - x_0\| \leq \delta_0^{q,x}$ with known $\delta_0^{q,x}$.

Assumption 7.3 (*Aleatoric + Epistemic Uncertainty*) The system is perturbed by random and bounded process and measurement noise signals, where the corresponding process noise w_k^q and measurement noise v_k^q are mutually uncorrelated, zero-mean "truncated" Gaussian white random signals with known covariance matrices, $Q_k^q = \mathbb{E}[w_k^q w_k^{q\top}] \succeq 0$ and $R_k^q = \mathbb{E}[v_k^q v_k^{q\top}] \succ 0$, and bounded norms, i.e., $\|w_k^q\| \leq \eta_w^q$ and $\|v_k^q\| \leq \eta_v^q$, respectively, where η_w^q and η_v^q are known. Moreover, x_0 is independent of v_k^q and w_k^q for all k, and an estimate $\hat{x}_0$ of the initial state x_0 is available, where $\|\hat{x}_0 - x_0\| \leq \delta_0^{q,x}$ with known $\delta_0^{q,x}$.

In the case of the stochastic/aleatoric uncertainty (i.e., if Assumption 7.1 holds and consequently, the uncertainty is characterized using probability distributions), the emphasis is on *expected/average* performance of the resilient state estimator. In this case, CPS safety/resilience is guaranteed based on probability of violation/chance constraints. On the other hand, in the case of set-valued/epistemic uncertainty (i.e., if Assumption 7.2 holds and hence the uncertainty is characterized by sets), the emphasis would be on the *best worst-case* performance and the CPS safety/resilience is

[1] Note that while the unknown inputs may also be used to represent uncertainties or noise that are unbounded or have unknown bounds, we primarily use this term to represent attack signals in this chapter and thus, we often use the terms unknown inputs and attacks interchangeably.

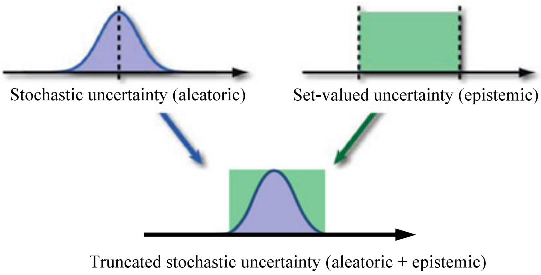
Stochastic uncertainty (aleatoric)
Set-valued uncertainty (epistemic)
Truncated stochastic uncertainty (aleatoric + epistemic)

Fig. 7.1 Different assumptions on the considered uncertainty in System (7.1)

guaranteed in the worst case, including rare events/corner cases. Finally, if Assumption 7.3 holds, we can combine the information of the stochastic uncertainties and the set-membership uncertainties from Assumptions 7.1 and 7.2 to benefit from the advantages of both. Figure 7.1 illustrates the aforementioned system uncertainty models/assumptions.

Both *categorical* and *continuous* natures of the uncertainties introduced by the switching and data injection attacks to the system of interest can be captured by the Cyber-Physical System (CPS) model in (7.1). The categorical nature of the switching and data injection attacks (*mode attack* and *signal location attack*) is modeled using the *hidden mode*, whereas the unknown input captures the continuous nature of the *signal magnitude attacks*. At any particular time k, the stochastic/bounded-error CPS is in precisely one of its modes, which is not measured, hence *hidden*.

Similar to Yong et al. (2018), we consider the model set $\mathcal{Q} \triangleq \mathcal{Q}^m \times \mathcal{Q}^d$ (whose cardinality will be characterized in Theorem 7.2 in Sect. 7.3.3.2) that include

(i) the modes of operation, $\mathcal{Q}^m$ (representing attacked switching mechanisms (e.g., circuit breakers, relays) via access to the jump set $\mathcal{D}_q$ and the mode transition function $\delta^q(\cdot)$, or the possible interconnection network topologies that affect the system matrices, A_k^q and B_k^q, and the sensor data network topologies, C_k^q and D_k^q) that an attacker can choose (*mode attack*), as well as

(ii) the different hypotheses for each mode, $\mathcal{Q}^d$, about which actuators and sensors are attacked or not attacked, represented by G_k^q and H_k^q, where our approach specifies which actuators and sensors are *not attacked*, in contrast to the approach in Mishra et al. (2015), which removes *attacked* sensor measurements and is not applicable for actuator attacks. (*signal location attack*).

More precisely, for sparse false data injection attacks, we let $G_k^q \triangleq \mathcal{G}_k \mathcal{I}_G^q$ and $H_k^q \triangleq \mathcal{H}_k \mathcal{I}_H^q$ for some input matrices $\mathcal{G}_k \in \mathbb{R}^{n \times t_a}$ and $\mathcal{H}_k \in \mathbb{R}^{\ell \times t_s}$, where t_a and t_s are the number of actuator and sensor signals that are *vulnerable*, respectively and encode the sparsity using $\mathcal{I}_G^q \in \mathbb{R}^{t_a \times p}$ and $\mathcal{I}_H^q \in \mathbb{R}^{t_s \times p}$ as index matrices such that $d_k^{a,q} \triangleq \mathcal{I}_G^q d_k$ and $d_k^{s,q} \triangleq \mathcal{I}_H^q d_k$ are subvectors of $d_k \in \mathbb{R}^p$ representing *signal magnitude attacks* on the actuators and sensors, respectively. These matrices provide a means to incorporate information about how the attacks affect the system, e.g., if the same attack is injected to an actuator and a sensor, or if some signals are *not* attacked, according to a particular hypothesis/mode q about the signal attack location.

The following are some examples from Yong et al. (2018) for choosing $\mathscr{G}_k$, $\mathscr{H}_k$, $\mathscr{I}_G^q$, and $\mathscr{I}_H^q$ to encode additional information about the nature/structure of data injection attacks.

Example 7.1 For a two-state system with two vulnerable actuators and one vulnerable sensor, if the same attack signal is injected into the first actuator and the sensor under the hypothesis corresponding to mode q, then $\mathscr{G}_k = I_2$, $\mathscr{H}_k = 1$, $\mathscr{I}_G^q = I_2$ and $\mathscr{I}_H^q = \begin{bmatrix}1 & 0\end{bmatrix}$. In this case, we obtain $G_k^q = I_2$ and $H_k^q = \begin{bmatrix}1 & 0\end{bmatrix}$.

Example 7.2 For a three-state system with three actuators and two sensors, if the first actuator and the second sensor are not vulnerable and there are three attacks according to the hypothesis corresponding to mode q, then $\mathscr{G}_k = \begin{bmatrix}0 & 0\\ 1 & 0\\ 0 & 1\end{bmatrix}$, $\mathscr{H}_k = \begin{bmatrix}1\\ 0\end{bmatrix}$, $\mathscr{I}_G^q = \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\end{bmatrix}$ and $\mathscr{I}_H^q = \begin{bmatrix}0 & 0 & 1\end{bmatrix}$. In this case, we have $G_k^q = \begin{bmatrix}0 & 0 & 0\\ 1 & 0 & 0\\ 0 & 1 & 0\end{bmatrix}$ and $H_k^q = \begin{bmatrix}0 & 0 & 1\\ 0 & 0 & 0\end{bmatrix}$.

Note that we assume that $p_a^q \leq t_a \leq m$ (i.e., the number of *attacked* actuator signals p_a^q under mode/hypothesis q cannot exceed the number of *vulnerable* actuators and in turn cannot exceed the total number of actuators m_a) and $p_s^q \leq t_s \leq \ell$ (with p_s^q *attacked* sensors from t_s *vulnerable* sensors out of ℓ measurements). Moreover, we assume that the maximum total number of attacks is $p \triangleq p_a^q + p_s^q \leq p^*$, where p^* is the maximum number of asymptotically correctable signal attacks (cf. Theorem 7.1 for its characterization).

7.2.2.1 System Assumptions

We require that the system is *strongly detectable*[2] in each mode. In fact, strong detectability is *necessary* for each mode in order to asymptotically correct the unknown attack signals, as shown in Yong et al. (2018) [Theorem 4.3] and is also necessary for deterministic systems [Sundaram and Hadjicostis (2007), Theorem 6]. Note that similar to the detectability property, strongly detectable systems need not be stable (cf. example in the proof of Theorem 7.1), but rather that the strongly undetectable modes of such systems are stable.

7.2.2.2 Knowledge of the System Operator/Defender

The matrices A_k^q, B_k^q, G_k^q, C_k^q, D_k^q, and H_k^q are known and the system (A_k^q, G_k^q, C_k^q, H_k^q) is strongly detectable in each mode. Further, the defender only knows (i) the

[2] A linear system is *strongly detectable* if $y_k = 0\ \forall k \geq 0$ implies $x_k \to 0$ as $k \to \infty$ for all initial states x_0 and input sequences $\{d_i\}_{i\in\mathbb{N}}$ (see [Yong et al. (2016b), Sect. 3.2] for necessary and sufficient conditions for this property).

upper bound on the *number* of actuators/sensors that can be attacked, p, and (ii) the switching mechanisms/topologies that may be compromised. The upper bound p allows the defender, in the worst case, to enumerate all possible combinations of G_k^q and H_k^q, while the latter assumption allows the defender to consider all possible topologies/modes of operations, representing A_k^q, B_k^q, C_k^q and D_k^q.

In addition, note that the above assumption of strong detectability can be viewed as recommendations or guidelines for system designers/operators to secure their systems as a preventative attack mitigation measure, since without strong detectability, resilient (i.e., unbiased or bounded) state estimates cannot be guaranteed. In other words, the requirement of strong detectability allows system designers to determine which actuators or sensors need to be safeguarded to guarantee resilient estimation.

7.2.3 Security Problem Statement

With the above modeling framework, the resilient state estimation problem can be posed as a problem of mode, state and input estimation, where the unknown inputs represent the unknown signal magnitude attacks and each mode/model represents an *attack mode* (resulting from the unknown mode attacks and unknown signal attack locations). The *objective* of this chapter is:

Problem 7.1 *Given an uncertain Cyber-Physical System (CPS) described by* (7.1)*,*

1. *Design a* resilient estimator *that asymptotically recovers* unbiased *estimates of the system state and attack signal in the presence of aleatoric/stochastic uncertainty (i.e., if Assumption 7.1 holds), or finds the set-valued estimates of compatible states and unknown inputs in the presence of epistemic uncertainty (i.e., if Assumption 7.2 holds), irrespective of the location or magnitude of attacks on its actuators and sensors as well as switching mechanism/topology (mode) attacks.*
2. *Investigate the fundamental limitations of the estimation algorithms, specifically the maximum number of asymptotically correctable signal attacks and the maximum number of required models with our multiple-model approach.*
3. *Find the conditions under which attacks can be detected and under which the attack strategy can be identified.*
4. *Design attack mitigation tools via $\mathscr{H}_\infty$-control with attack rejection.*

7.3 Resilient State Estimation

Similar to a previous approach for stochastic systems in Yong et al. (2021), we propose the use of a *multiple-model* estimation approach to solve Problem 7.1.1. Then, we will consider Problem 7.1.2 and characterize some fundamental limitations to resilient estimation in Sect. 7.3.3.

7.3.1 Multiple-Model State and Input Filtering/Estimation Algorithm

Inspired by the multiple-model filtering algorithms for hidden-mode hybrid systems with *known* inputs (e.g., Bar-Shalom et al. (2004); Mazor et al. (1998) and references therein), our multiple-model (MM) framework (see Fig. 7.2) consists of three components: (i) a bank of mode-matched filters/observers, (ii) a mode estimator that finds the most likely or compatible modes, and (iii) a global fusion estimator that combines/fuses states and unknown input (attack) estimates from (i) based on the estimated modes in (ii), which are described in greater detail below.

7.3.1.1 Mode-Matched Filters/Observers

The bank of filters/observers is comprised of $\mathfrak{N}$ simultaneous state and input filters/observers, one for each mode, that differ based on the assumptions on system uncertainties and noise signals. If Assumption 7.1 (the aleatoric/stochastic uncertainty model) holds, the optimal recursive filter developed in Yong et al. (2016b) can be applied, while if Assumption 7.2 (the epistemic/set-valued uncertainty model) holds, the recursive set-valued observer developed in Yong (2018) can be utilized. Both variants are recursive and involve the same three-step structure as follows:

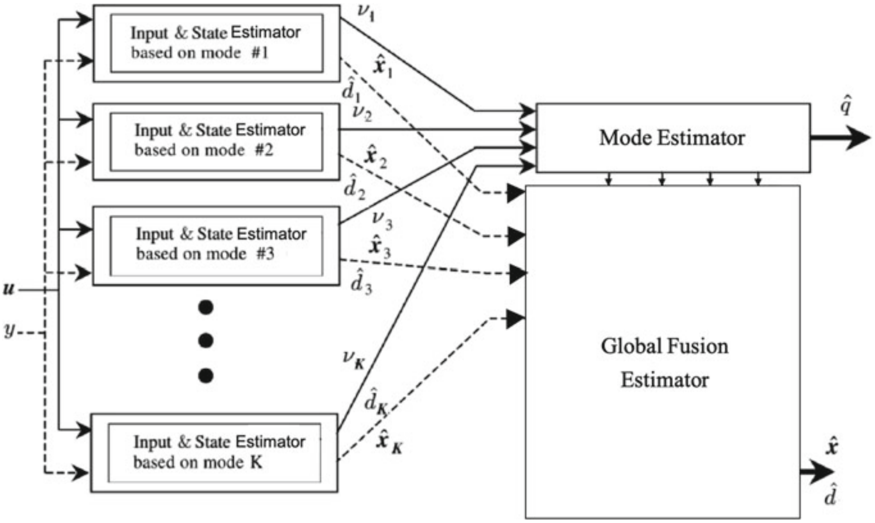

Fig. 7.2 Multiple-model framework for hidden mode, input and state estimation, which consists of a (i) bank of mode-matched filters/observers, (ii) a mode estimator and (iii) a global fusion estimator

Unknown Input Estimation:

$$\begin{aligned}\hat{d}^q_{1,k} &= M^q_{1,k}(z^q_{1,k} - C^q_{1,k}\hat{x}^q_{k|k} - D^q_{1,k}u^q_k),\\ \hat{d}^q_{2,k-1} &= M^q_{2,k}(z^q_{2,k} - C^q_{2,k}\hat{x}^q_{k|k-1} - D^q_{2,k}u^q_k),\\ \hat{d}^q_{k-1} &= V^q_{1,k-1}\hat{d}^q_{1,k-1} + V^q_{2,k-1}\hat{d}^q_{2,k-1}.\end{aligned} \tag{7.2}$$

Time Update:

$$\begin{aligned}\hat{x}^q_{k|k-1} &= A^q_{k-1}\hat{x}^q_{k-1|k-1} + B^q_{k-1}u^q_{k-1} + G^q_{1,k-1}\hat{d}^q_{1,k-1},\\ \hat{x}^{\star,q}_{k|k} &= \hat{x}^q_{k|k-1} + G^q_{2,k-1}\hat{d}^q_{2,k-1}.\end{aligned} \tag{7.3}$$

Measurement Update:

$$\hat{x}^q_{k|k} = \hat{x}^{\star,q}_{k|k} + \tilde{L}^q_k(z^q_{2,k} - C^q_{2,k}\hat{x}^{\star,q}_{k|k} - D^q_{2,k}u^q_k), \tag{7.4}$$

where $\hat{x}^q_{k-1|k-1}$, $\hat{d}^q_{1,k-1}$, $\hat{d}^q_{2,k-1}$ and $\hat{d}^q_{k-1}$ denote the optimal *point* estimates of x^q_{k-1}, $d^q_{1,k-1}$, $d^q_{2,k-1}$ and d^q_{k-1}, respectively, if Assumption 7.1 holds (cf. Algorithm 7.1 that summarizes the optimal filter for mode q in the presence of stochastic (aleatoric) uncertainty) and denote the centroids of the *hyperball-valued* estimates of x^q_{k-1}, $d^q_{1,k-1}$, $d^q_{2,k-1}$ and d^q_{k-1}, respectively, if Assumption 7.2 holds (cf. Algorithm 7.3 that finds the $\mathscr{H}_\infty$-optimal set-valued state and input estimates for mode q in the presence of distribution-free (epistemic) uncertainty).

The rest of the notations are clarified in the context of the system transformation described in Appendix 7.1.1. For details of the filter/observer derivation of both variants, as well as necessary and sufficient conditions for filter stability and optimality of the mode-matched filters/observers, the reader is referred to Yong et al. (2016b) and Yong (2018) for the aleatoric and epistemic uncertainty models, respectively.

It is worth mentioning that in the case that Assumption 7.3 holds (i.e., with a combination of aleatoric and epistemic uncertainties), we can compute (in parallel) both the point estimates corresponding to aleatoric/stochastic uncertainty and the set-valued estimates corresponding to the epistemic/bounded-error uncertainty, and utilize their combination as described in the following subsections.

7.3.1.2 Mode Estimator

The mode estimator seeks to determine the most likely or all compatible modes based on the observations. For this purpose, we consider three cases:

(a) **Aleatoric Uncertainty.** In this case, Assumption 7.1 holds and consequently, a *mode probability computation* is performed for all modes as described in Yong et al. (2018). The multiple-model approach computes the probability of each mode by exploiting the whiteness property [Yong et al. (2021), Theorem 1] of the *generalized innovation* sequence, ν^q_k, defined as

$$\nu^q_k \triangleq \tilde{\Gamma}^q_k(z^q_{a,2,k} - C^q_{a,2,k}\hat{x}^{\star,q}_{a,k|k} - D^q_{a,2,k}u^q_k), \tag{7.5}$$

i.e., $\nu_k^q \sim \mathcal{N}(0, S_k^q)$ (a multivariate normal distribution) with covariance $S_k^q \triangleq \mathbb{E}[\nu_k^q \nu_k^{q\top}] = \tilde{\Gamma}_k^q \tilde{R}_{2,k}^{\star,q} \tilde{\Gamma}_k^{q\top}$ and where $\tilde{\Gamma}_k^q$ is chosen such that S_k^q is invertible and $\tilde{R}_{2,k}^{\star,q}$ is given in Algorithm 7.1. This generalized innovation represents a residual signal with false data injection attacks removed that can be used to define the *likelihood function* for each mode q at time k conditioned on all prior measurements Z^{k-1}:

$$\mathscr{L}(q|z_{2,k}) \triangleq \mathcal{N}(\nu_k^q; 0, S_k^q) = \frac{\exp(-\frac{1}{2}\nu_k^{q\top}(S_k^q)^{-1}\nu_k^q)}{\sqrt{|2\pi S_k^q|}}. \tag{7.6}$$

Then, the posterior probability μ_k^j for each mode j is recursively computed from the prior probability μ_{k-1}^j using Bayes' rule as follows:

$$\mu_k^j = P(q = j|z_{1,k}, z_{2,k}, Z^{k-1}) = \frac{\mathcal{N}(\nu_k^j; 0, S_k^j)\mu_{k-1}^j}{\sum_{i=1}^{\mathfrak{N}} \mathcal{N}(\nu_k^i; 0, S_k^i)\mu_{k-1}^j}. \tag{7.7}$$

Furthermore, to keep the modes "alive" in case of a switch in the attacker's strategy, a heuristic lower bound on all mode probabilities is imposed.

(b) **Epistemic Uncertainty.** In the presence of distribution-free and bounded norm noise signals, i.e., when Assumption 7.2 holds, a *mode elimination* process is performed to eliminate the modes that are incompatible with observations, which results in a set of compatible modes. The mode elimination approach relies on the checking of some *residual* signals against some thresholds. We first define the residual signal r_k^q for each mode q at time step k as:

$$r_k^q \triangleq z_{e,2,k}^q - C_{e,2,k}^q \hat{x}_{a,k|k}^{\star,q} - D_{e,2,k}^q u_k^q. \tag{7.8}$$

Then, leveraging an approach in Khajenejad and Yong (2019), if the residual signal of a particular mode exceeds its upper bound conditioned on this mode being true, we can conclusively rule it out as incompatible. To do so, for each mode q, we compute a *tractable* upper bound ($\hat{\delta}_{r,k}^q$; cf. Proposition 7.2) for the 2-norm of its corresponding residual at time k, conditioned on q being the *true* mode. Then, comparing the 2-norm of residual signal in (7.8) with $\hat{\delta}_{r,k}^q$, we can eliminate mode q if the residual's 2-norm is strictly greater than the upper bound, i.e., if $\|r_k^q\|_2 > \hat{\delta}_{r,k}^q$. This can be formalized using the following proposition (cf. [Khajenejad and Yong (2019), Proposition 1 and Theorem 2] for more details and a formal proof of this result).

Proposition 7.1 *Consider mode q and its residual signal r_k^q at time step k. Assume that $\delta_{r,k}^{q,*}$ is any signal that satisfies $\|r_k^{q|*}\|_2 \leq \delta_{r,k}^{q,*}$, where $r_k^{q|*}$ is the true mode's residual signal (i.e., $q = q^*$, where q^* denotes the true mode), defined as follows:*

$$r_k^{q|*} \triangleq z_{e,2,k}^{q*} - C_{e,k,2}^q \hat{x}_{e,k|k}^{\star,q} - D_{e,k,2}^q u_k^q = T_{e,k,2}^{q*} y_k - C_{e,k,2}^q \hat{x}_{e,k|k}^{\star,q} - D_{e,k,2}^q u_k^q. \tag{7.9}$$

Then, mode q is not the true mode, i.e., can be eliminated at time k, if

$$\|r^q_k\|_2 > \delta^{q,*}_{r,k}. \tag{7.10}$$

Note that by [Khajenejad and Yong (2019), Lemmas 1 and 2], the sequence $\{\delta^{q,*}_{r,k}\}^{\infty}_{k=0}$ is uniformly bounded and admits a finite valued upper sequence. Although computing the tightest possible residual norm's upper sequence potentially can eliminate the most possible number of modes, it requires to the solution a *norm maximization* problem over the intersection of level sets of lower dimensional norm functions that is NP-hard [Bodlaender et al. (1990)]. Thus, by applying [Khajenejad and Yong (2019), Theorem 3], we instead compute a tractable over-approximation of the residual norm's upper bound sequence, denoted by $\{\hat{\delta}^q_{r,k}\}^{\infty}_{k=0}$, i.e., $\forall k \in \{0, \ldots, \infty\}$, $\delta^{q,*}_{r,k} \leq \hat{\delta}^q_{r,k}$, and use this upper bound sequence as a tractable mode elimination criterion as follows (cf. [Khajenejad and Yong (2019), Theorem 3] for more details):

Proposition 7.2 *Mode q is not the true mode, i.e., can be eliminated at time k, if*

$$\|r^q_k\|_2 > \hat{\delta}^q_{r,k} \triangleq \min\{\delta^{q,inf}_{r,k}, \delta^{q,tri}_{r,k}\}, \tag{7.11}$$

where $\delta^{q,inf}_{r,k}$ and $\delta^{q,tri}_{r,k}$ are two tractable computed upper bounds for the residual norm and are given in Appendix 7.1.2.

(c) **Combined Uncertainty.** In the presence of truncated Gaussian noise signals, i.e., if Assumption 7.3 holds, both mode probability computation procedure (described in (7.3.1.2)) and mode elimination approach (described in (7.3.1.2)) are applicable and can be combined. Specifically, we first apply the mode elimination algorithm from Khajenejad and Yong (2019) to obtain a set of compatible modes, and then compute mode probabilities for only the "non-eliminated" modes using (7.7).

7.3.1.3 Global Fusion Estimator

Finally, the global fusion estimator combines the estimates from the bank of mode-matched state and input estimators and mode observer, under the three different system uncertainty models, as follows:

(a) **Aleatoric Uncertainty.** Based on the posterior mode probabilities in (7.7), the most likely mode at each time k, $\hat{q}_k$, and the associated state and input estimates and covariances, $\hat{x}_{a,k|k}$, $\hat{d}_{a,k}$, $P^x_{k|k}$ and P^d_k, can be determined:

$$\begin{aligned} \hat{q}_k &= j^* = \arg\max_{j\in\{1,\ldots,\mathfrak{N}\}} \mu^j_k, \\ \hat{x}_{a,k|k} &= \hat{x}^{j^*}_{a,k|k}, \quad \hat{d}_{a,k} = \hat{d}^{j^*}_{a,k}, \\ P^x_{k|k} &= P^{x,j^*}_{k|k}, \quad P^d_k = P^{d,j^*}_k. \end{aligned} \tag{7.12}$$

(b) **Epistemic Uncertainty.** Using the computed residuals (7.9) and their upper bound sequences (7.11), our proposed global fusion observer finds all modes that are not eliminated and computes the input and state set-valued estimates, $\hat{D}_{k-1}$ and $\hat{X}_k$, by taking the union of the mode-matched state and unknown input (attack) set estimates over the compatible modes:

$$\begin{aligned} \hat{\mathscr{Q}}_k &= \{q \in \mathscr{Q} \mid \|r^q\|_2 \leq \hat{\delta}^q_{r,k}\}, \\ \hat{D}_{k-1} &= \cup_{q \in \hat{\mathscr{Q}}_k} D^q_{k-1}, \\ \hat{X}_k &= \cup_{q \in \hat{\mathscr{Q}}_k} X^q_k. \end{aligned} \tag{7.13}$$

(c) **Combined Uncertainty.** In this case, after eliminating all modes that satisfy (7.11), the most likely mode and its associated state and input estimates and covariances at each time can be determined using only the set of non-eliminated modes (instead of all modes as in the case of aleatoric uncertainty), i.e.,

$$\begin{aligned} \hat{\hat{q}}_k \quad &= j^{**} = \arg\max_{j \in \hat{\mathscr{Q}}} \mu^j_k, \\ \hat{x}_{c,k|k} &= \hat{x}^{j^{**}}_{c,k|k}, \quad \hat{d}_{c,k} = \hat{d}^{j^{**}}_{c,k}, \\ P^x_{c,k|k} &= P^{x,j^{**}}_{k|k}, \quad P^d_{c,k} = P^{d,j^{**}}_k. \end{aligned} \tag{7.14}$$

The multiple-model approach is summarized in Algorithms 7.1–7.4 for the aleatoric/stochastic and epistemic/set-valued uncertainties, respectively.

7.3.2 *Properties of the Resilient State Estimator*

Our previous results in Yong et al. (2021); Khajenejad and Yong (2019); Yong (2018) show that the resilient state estimator has nice properties, which can be summarized as follows.

7.3.2.1 Optimality

Given the attacked switched linear system with hidden modes in (7.1), if Assumption 7.1 holds (aleatoric uncertainty), the resilient state estimator (i.e., Algorithms 7.1 and 7.2) is *asymptotically optimal*, i.e., the state and input estimates in (7.12) converge on average to optimal state and input estimates in the minimum variance unbiased sense [Yong et al. (2021), Corollary 13]. On the other hand, if Assumption 7.2 holds (epistemic uncertainty), the resulting set-valued estimates in (7.13) are uniformly bounded [Yong (2018), Lemma 1] and the resilient state and input observer is stable and optimal in the $\mathscr{H}_\infty$-norm sense [Yong (2018) [Theorem 2]]. Further, in the presence of truncated Gaussian noise signals, i.e., if Assumption 7.3 is satisfied, it can be shown that the set-valued estimates are uniformly bounded, but the resilient state estimates obtained from Algorithms 7.3 and 7.4, may not be asymptotically optimal.

7.3.2.2 Mode Detectability

Given the attacked switched linear system with hidden modes in (7.1), in the presence of aleatoric/stochastic uncertainty, i.e., if Assumption 7.1 holds, the resilient state estimator is *mean consistent*, i.e., the geometric mean of the mode probability for the true model $q^* \in \mathcal{Q}$ asymptotically converges to one for all initial mode prob-

Algorithm 7.1 OPT- FILTER finds the optimal state and input estimates for mode q in the presence of stochastic (aleatoric) uncertainty

```
Input: $q, \hat{x}^q_{k-1|k-1}, \hat{d}^q_{1,k-1}, P^{x,q}_{k-1|k-1}, P^{xd,q}_{1,k-1}, P^{d,q}_{1,k-1}$
[superscript “$q$" and subscript “$a$" (referring to aleatoric uncertainty) omitted in the following]
▷ Estimation of $d_{2,k-1}$ and $d_{k-1}$
$\hat{A}_{k-1} = A_{k-1} - G_{1,k-1} M_{1,k-1} C_{1,k-1};$
$\hat{Q}_{k-1} = G_{1,k-1} M_{1,k-1} R_{1,k-1} M^\top_{1,k-1} G^\top_{1,k-1} + Q_{k-1};$
$\tilde{P}_k = \hat{A}_{k-1} P^x_{k-1|k-1} \hat{A}^\top_{k-1} + \hat{Q}_{k-1};$
$\tilde{R}_{2,k} = C_{2,k} \tilde{P}_k C^\top_{2,k} + R_{2,k};$
$P^d_{2,k-1} = (G^\top_{2,k-1} C^\top_{2,k} \tilde{R}^{-1}_{2,k} C_{2,k} G_{2,k-1})^{-1};$
$M_{2,k} = P^d_{2,k-1} G^\top_{2,k-1} C^\top_{2,k} \tilde{R}^{-1}_{2,k};$
$\hat{x}_{k|k-1} = A_{k-1} \hat{x}_{k-1|k-1} + B_{k-1} u_{k-1} + G_{1,k-1} \hat{d}_{1,k-1};$
$\hat{d}_{2,k-1} = M_{2,k}(z_{2,k} - C_{2,k} \hat{x}_{k|k-1} - D_{2,k} u_k);$
$\hat{d}_{k-1} = V_{1,k-1} \hat{d}_{1,k-1} + V_{2,k-1} \hat{d}_{2,k-1};$
$P^d_{12,k-1} = M_{1,k-1} C_{1,k-1} P^x_{k-1|k-1} A^\top_{k-1} C^\top_{2,k} M^\top_{2,k} - P^d_{1,k-1} G^\top_{1,k-1} C^\top_{2,k} M^\top_{2,k};$
$P^d_{k-1} = V_{k-1} \begin{bmatrix} P^d_{1,k-1} & P^d_{12,k-1} \\ P^{d\top}_{12,k-1} & P^d_{2,k-1} \end{bmatrix} V^\top_{k-1};$
▷ Time update
$\hat{x}^\star_{k|k} = \hat{x}_{k|k-1} + G_{2,k-1} \hat{d}_{2,k-1};$
$P^{\star x}_{k|k} = G_{2,k-1} M_{2,k} R_{2,k} M^\top_{2,k} G^\top_{2,k} + (I - G_{2,k-1} M_{2,k} C_{2,k}) \tilde{P}_k (I - G_{2,k-1} M_{2,k} C_{2,k})^\top;$
$\tilde{R}^\star_{2,k} = C_{2,k} P^{\star x}_{k|k} C^\top_{2,k} + R_{2,k} - C_{2,k} G_{2,k-1} M_{2,k} R_{2,k} - R_{2,k} M^\top_{2,k} G^\top_{2,k-1} C_{2,k};$
▷ Measurement update
$\breve{P}_k = P^{\star x}_{k|k} C^\top_{2,k} - G_{2,k-1} M_{2,k} R_{2,k};$
$\tilde{L}_k = \breve{P}_k \tilde{R}^{\star\dagger}_{2,k};$
$\hat{x}_{k|k} = \hat{x}^\star_{k|k} + \tilde{L}_k (z_{2,k} - C_{2,k} \hat{x}^\star_{k|k} - D_{2,k} u_k);$
$P^x_{k|k} = \tilde{L}_k R^\star_{2,k} \tilde{L}^\top_k - \tilde{L}_k \breve{P}^\top_k - \breve{P}_k \tilde{L}^\top_k;$
▷ Estimation of $d_{1,k}$
$\tilde{R}_{1,k} = C_{1,k} P^x_{k|k} C^\top_{1,k} + R_{1,k};$
$M_{1,k} = \Sigma^{-1}_k;$
$P^d_{1,k} = M_{1,k} \tilde{R}_{1,k} M_{1,k};$
$\hat{d}_{1,k} = M_{1,k}(z_{1,k} - C_{1,k} \hat{x}_{k|k} - D_{1,k} u_k);$
return $\tilde{R}^{\star,q}_{2,k}, \hat{x}^{\star,q}_{k|k}$
```

Algorithm 7.2 RESILIENT STATE ESTIMATOR (STATIC- MM- ESTIMATOR) finds resilient state estimates corresponding to most likely mode in the presence of stochastic (aleatoric) uncertainty

```
Input: $\forall j \in \{1, 2, \ldots, \mathfrak{N}\}$: $\hat{x}^j_{0|0}$; $\mu^j_0$;
[subscript "$a$" (referring to aleatoric uncertainty) omitted in the following]
$\hat{d}^j_{1,0} = (\Sigma^j_0)^{-1}(z^j_{1,0} - C^j_{1,0}\hat{x}^j_{0|0} - D^j_{1,0}u_0)$;
$P^{d,j}_{1,0} = (\Sigma^j_0)^{-1}(C^j_{1,0}P^{x,j}_{0|0}C^{j\top}_{1,0} + R^j_{1,0})(\Sigma^j_0)^{-1}$;
for $k = 1$ to $N$ do
    for $j = 1$ to $\mathfrak{N}$ do
        ▷ Mode-Matched Filtering  Run OPT- FILTER($j$,$\hat{x}^j_{k-1|k-1}$, $\hat{d}^j_{1,k-1}$, $P^{x,j}_{k-1|k-1}$, $P^{d,j}_{1,k-1}$);
        $\overline{\nu}^j_k \triangleq z^j_{2,k} - C^j_{2,k}\hat{x}^{\star,j}_{k|k} - D^j_{2,k}u_k$;
        $\mathscr{L}(j|z^j_{2,k}) = \frac{1}{(2\pi)^{p^j_{\tilde{R}}/2}|\tilde{R}^{j,\star}_{2,k}|^{1/2}_+} \exp\left(-\frac{\overline{\nu}^{j\top}_k \tilde{R}^{j,\star\dagger}_{2,k}\overline{\nu}^j_k}{2}\right)$;
    for $j = 1$ to $\mathfrak{N}$ do
        ▷ Mode Probability Update (small $\epsilon > 0$)
        $\overline{\mu}^j_k = \max\{\mathscr{L}(j|z^j_{2,k})\mu^j_{k-1}, \epsilon\}$;
    for $j = 1$ to $\mathfrak{N}$ do
        ▷ Mode Probability Update (normalization)
        $\mu^j_k = \frac{\overline{\mu}^j_k}{\sum_{\ell=1}^{\mathfrak{N}} \overline{\mu}^\ell_k}$;
        ▷ Output
        Compute (7.12);
return $\hat{x}_{k|k}$, $P^x_{k|k}$
```

abilities [Yong et al. (2021), Theorem 8]. Furthermore, in the case of epistemic/set-valued uncertainty, i.e., if Assumption 7.2 holds, the resilient state estimator is *mode detectable* by [Khajenejad and Yong (2019), Theorem 4], i.e., there exists a natural number $K > 0$, such that for all time steps $k \geq K$, all false modes are eliminated, if either the whole observation/measurement and state spaces are bounded or the unknown input/attack signal has an *unlimited energy*, as well as some additional mild conditions hold (cf. [Khajenejad and Yong (2019), Assumptions 1&2, Lemmas 3–5 and Theorem 4] for more details). Similarly, if Assumption 7.3 holds, all false modes (except for the true mode) will be eliminated after some large enough finite time under the same assumption of bounded state spaces or unlimited energy, and the unique true mode will have probability one.

7.3.3 Fundamental Limitations of Attack-Resilient Estimation

Next, to address Problem 1.2, we characterize fundamental limitations of the attack-resilient estimation problem and of our multiple mode filtering/estimation approach.

Algorithm 7.3 Opt- Observer finds the $\mathscr{H}_\infty$-optimal set-valued state and input estimates for mode q in the presence of distribution-free (epistemic) uncertainty

```
**Input**: $q, \hat{x}^q_{k-1|k-1}, \hat{d}^q_{k-1}$
[superscript "$q$" and subscript "$e$" (referring to the epistemic (set-valued) uncertainty) omitted in the following]
▹ Estimation of $d_{2,k-1}$ and $d_{k-1}$
$M_{1,k} = \Sigma_k^{-1}$,
$M_{2,k} = (C_{2,k}G_{2,k})^\dagger$,
$\hat{A}_{k-1} = A_{k-1} - G_{1,k-1}M_{1,k-1}C_{1,k-1}$;
$\Phi_k = I - G_{2,k}M_{2,k}C_{2,k}$;
$\overline{A}_k = \Phi_k\hat{A}_k$;
$V_{e,k} = V_{1,k}M_{1,k}C_{1,k} + V_{2,k}M_{2,k}C_{2,k}\hat{A}_k$;
$A_{e,k} = (I - \tilde{L}_kC_{2,k})\overline{A}_k$;
$B_{e,w,k} = (I - \tilde{L}_kC_{2,k})\Phi_k$;
$B_{e,v_1,k} = -(I - \tilde{L}_kC_{2,k})\Phi_kG_{1,k}M_{1,k}T_{1,k}$;
$B_{e,v_2,k} = -((I - \tilde{L}_kC_{2,k})G_{2,k}M_{2,k} + \tilde{L}_k)T_{2,k}$;
$\hat{x}_{k|k-1} = A_{k-1}\hat{x}_{k-1|k-1} + B_{k-1}u_{k-1} + G_{1,k-1}\hat{d}_{1,k-1}$;
$\hat{d}_{2,k-1} = M_{2,k}(z_{2,k} - C_{2,k}\hat{x}_{k|k-1} - D_{2,k}u_k)$;
$\hat{d}_{k-1} = V_{1,k-1}\hat{d}_{1,k-1} + V_{2,k-1}\hat{d}_{2,k-1}$;
$\delta^d_{k-1} = \delta^x_0\|V_{e,k}A^{k-1}_{e,k}\| + \eta_w(\sum_{i=0}^{k-2}\|V_{e,k}A^{k-2-i}_{e,k}B_{e,w,k}\| + \|V_{2,k}M_{2,k}C_{2,k}\|) +$
$\eta_v(\|V_{2,k}M_{2,k}T_{2,k}\| + \|V_{e,k}A^{k-2}_{e,k}B_{e,v_1,k}\| + \|V_{e,k}B_{e,v_2,k} + (V_{1,k} - V_{2,k}M_{2,k}C_{2,k}G_{1,k})M_{1,k}T_{1,k}\| +$
$\sum_{i=1}^{k-2}\|V_{e,k}A^{k-2-i}_{e,k}(B_{e,v_1,k} + A_{e,k}B_{e,v_2,k})\|)$;
$\hat{D}_{k-1} = \{d \in \mathbb{R}^l : \|d - \hat{d}_{k-1}\| \leq \delta^d_{k-1}\}$;
▹ Time update
$\hat{x}^\star_{k|k} = \hat{x}_{k|k-1} + G_{2,k-1}\hat{d}_{2,k-1}$;
▹ Measurement update
$\hat{x}_{k|k} = \hat{x}^\star_{k|k} + \tilde{L}_k(z_{2,k} - C_{2,k}\hat{x}^\star_{k|k} - D_{2,k}u_k)$;
$\delta^x_k = \delta^x_0\|A^k_{e,k}\| + \eta_w\sum_{i=0}^{k-1}\|A^i_{e,i}B_{e,w,i}\| + \eta_v(\|B_{e,v_2,k}\| + \|A^{k-1}_{e,k}B_{e,v_1,k}\| +$
$\sum_{i=0}^{k-2}\|A^i_{e,i}(B_{e,v_1,i} + A_{e,k}B_{e,v_2,i})\|)$;
$\hat{X}_k = \{x \in \mathbb{R}^n : \|x - \hat{x}_{k|k}\| \leq \delta^x_k\}$;
▹ Estimation of $d_{1,k}$
$\hat{d}_{1,k} = M_{1,k}(z_{1,k} - C_{1,k}\hat{x}_{k|k} - D_{1,k}u_k)$;
**return** $\hat{X}^q_k, \hat{D}^q_{k-1}$
```

Note that these fundamental limitations apply to all hidden-mode switched linear systems with unknown inputs (attacks) (7.1), regardless of the assumptions about the system uncertainties. First, under the assumption that there is only false data injection attacks (no switching attacks), we find an upper bound on the number of correctable signal attacks/errors (i.e., signal attacks whose effects can be negated or cancelled). Then, we characterize the maximum number of models that is required by our multiple-model approach to obtain resilient estimates despite attacks.

Algorithm 7.4 RESILIENT MODE, STATE AND INPUT ESTIMATOR simultaneously finds compatible sets of modes, unknown inputs (attacks) and states in the presence of distribution-free (epistemic) uncertainties

```
Input: $\mathscr{Q} \triangleq \{1, 2, \ldots, \mathfrak{N}\}$, $\forall j \in \{1, 2, \ldots, \mathfrak{N}\}$: $\hat{x}^j_{0|0}$;
[subscript "$e$" (referring to the epistemic (set-valued) uncertainty) omitted in the following]
$\hat{\mathscr{Q}}_0 = \mathscr{Q}$;
for $k = 1$ to $N$ do
    for $q \in \hat{\mathscr{Q}}_{k-1}$ do
        ▷Mode-Matched State and Input Set-Valued Estimates
        Run OPT- OBSERVER($q$,$\hat{x}^q_{k-1|k-1}$, $\hat{d}^q_{k-1}$);
        $z^q_{2,k} = T^q_2 y_k$;
        ▷Mode Observer via Elimination
        $\hat{\mathscr{Q}}_k = \hat{\mathscr{Q}}_{k-1}$;
        Compute $r^q$ via (7.8)
        and $\hat{\delta}^q_{r,k}$ via Proposition 7.2;
        if $\|r^q\|_2 > \hat{\delta}^q_{r,k}$ then
            $\hat{\mathscr{Q}}_k = \hat{\mathscr{Q}}_k \backslash \{q\}$;
    ▷State and Input Estimates
    $\hat{X}_k = \cup_{q \in \hat{\mathscr{Q}}_k} \hat{X}^q_k$;
    $\hat{D}_{k-1} = \cup_{q \in \hat{\mathscr{Q}}_k} \hat{D}^q_{k-1}$;
return $\hat{\mathscr{Q}}_k$, $\hat{D}_{k-1}$, $\hat{X}_k$
```

7.3.3.1 Number of Asymptotically Correctable Signal Attacks

We begin by defining the notion of correctable signal attacks in the setting with only data injection attacks, which is itself an interesting CPS security research problem.

Definition 7.1 (Correctable Signal Attacks) We say that p actuators and sensors signal attacks are correctable, if for any initial state $x_0 \in \mathbb{R}^n$ and signal attack sequence $\{d_j\}_{j \in \mathbb{N}}$ in $\mathbb{R}^p$, we have an estimator/observer such that the estimate bias asymptotically/exponentially tends to zero (under aleatoric unceraintiy, cf. Assumption 7.1), i.e., $\mathbb{E}[\hat{x}_{a,k|k} - x_k] \to 0$ (and $\mathbb{E}[\hat{d}_{a,k-1} - d_{k-1}] \to 0$) as $k \to \infty$ or if the set estimation errors are ultimately uniformly bounded sequences (under epistemic uncertainty, cf. Assumption 7.2).

To derive an estimation-theoretic upper bound on the maximum number of signal attacks that can be asymptotically corrected, we assume that the true model or mode ($q = q^*$) is known. Thus, depending on the type of uncertainty, the resilient state estimation problem is identical to the state and input estimation problem in Yong et al. (2016b) or Yong (2018), where the unknown inputs represent the attacks on the actuator and sensor signals. It has been shown in Yong et al. (2016b) and Yong (2018) that the system property of strong detectability is a necessary condition for obtaining uniformly bounded estimates (cf. Yong et al. (2016b); Yong (2018) for more details, e.g., regarding filter/observer stability and existence). Thus, we will use this necessary system property to find an upper bound on the maximum number of signal attacks that can be corrected, similar to Yong et al. (2018), as follows:

Theorem 7.1 (Maximum Correctable Data Injection Attacks) *The maximum number of correctable actuators and sensors signal attacks, p^*, for system* (7.1) *is equal to the number of sensors, l, i.e., $p^* \leq l$ and the upper bound is achievable.*

Proof A necessary and sufficient condition for strong detectability (with the true model $q = q^*$) is given in Yong et al. (2016b); Yong (2018) as

$$\text{rk}\begin{bmatrix} zI - A^* & -G^* \\ C^* & H^* \end{bmatrix} = n + p^*, \ \forall z \in \mathbb{C}, |z| \geq 1. \tag{7.15}$$

Since the above system matrix has only $n + l$ rows, it follows that its rank is at most $n + l$. Thus, from the necessary condition for (7.15), we obtain $n + p^* \leq n + l \Rightarrow p^* \leq l$. The upper bound is achievable using the example of the discrete-time equivalent model (with time step $\Delta t = 0.1s$) of the smart grid case study in Liu et al. (2013), as shown in Yong et al. (2018) [Theorem 4.3]. ■

The above result means that for each mode, the total number of vulnerable actuators and sensors must not exceed the number of measurements, which can serve as a guide for *preventative attack mitigation*, where the actuators or sensors that need to be safeguarded to guarantee resilient estimation can be determined. Note that the result in Theorem 7.1 is stronger than the standard and well-known result in the literature (e.g., in Fawzi et al. 2014, Proposition 3), where the maximum number of correctable attacks is at most equal to half of the number of sensors, presumably since we only require strong detectability instead of strong observability.

7.3.3.2 Number of Required Models for Estimation Resilience

Next, returning to the more general case with false date injection as well as switching attacks, i.e., the hidden-mode switched linear system in (7.1), we characterize the maximum number of models $\mathfrak{N}^*$ that are needed with the multiple-model approach in Sect. 7.3.1, which is independent of the size of the system, e.g., the number of buses in a power system, as well as the type/model of system uncertainty:

Theorem 7.2 (Maximum Number of Models/Modes) *Suppose there are t_a actuators and t_s sensors, and at most $p \leq l$ of these signals are attacked. Suppose also that there are t_m possible attack modes (*mode attack*). Then, the combinatorial number of all possible models, and hence the maximum number of models that need to be considered with the multiple-model approach, is*

$$\mathfrak{N}^* = t_m \binom{t_a + t_s}{p} = t_m \binom{t_a + t_s}{t_a + t_s - p}.$$

Proof It is sufficient to consider only models corresponding to the maximum number of attacks p. All models with strictly less than p attacks are contained in this set of models with the attack vectors having some identically zero elements for which our estimation algorithm is still applicable. Thus, we only need to consider combinations of p attacks among $t_a + t_s$ sensors and actuators for each of the t_m attack modes of operation/topologies. Note that this number is the maximum because resilience may be achievable with less models: For instance, when $t_m = 1$, $t_a = 0$ and $t_s = 2 = l$, $p = 1$, $A = \begin{bmatrix} 0.1 & 1 \\ 0 & 0.2 \end{bmatrix}$ and $C = I_2$, we have $\mathfrak{N}^* = 2$, but it can be verified that with $G = 0_{2\times2}$ and $H = I_2$ (only one model, i.e., $1 = \mathfrak{N} < \mathfrak{N}^*$), the system is strongly detectable. ■

Note that the number of required models may change if additional knowledge about the data injection attack strategies is available. For instance, if we know that there are at most $n_a \leq t_a$ and $n_s \leq t_s$ attacks on the actuators and sensors, respectively, with a total of p attacks (where $p \leq l$ and $p \leq n_a + n_s$), then the maximum number of models that are required,

$$\mathfrak{N}^* = t_m \sum_{i=0}^{\min\{n_a, p\}} \binom{t_a}{i} \binom{t_s}{\min\{p - i, n_s\}}$$

is less than the number required in combinatorial case in Theorem 7.2.

On the other hand, the number of models may actually increase with less vulnerable actuators or sensors, as shown in the following example with $t_m = 1$ (one mode of operation), $n_a = 0$ (no attacks on actuators), $A = \begin{bmatrix} 0.1 & 1 \\ 0 & 1.2 \end{bmatrix}$ and $C = I$. If only one of the two sensors is vulnerable ($n_s = p = 1 < l = 2$), we have two models with $G = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $H_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $H_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, but if both sensors are vulnerable ($n_s = p = 2$), only one model is required with $G = 0$ and $H = I$. Note that the latter case is not strongly detectable with zeros at $\{0.1, 1.2\}$, thus this system violates the necessary condition in Yong et al. (2016b); Yong (2018) for obtaining resilient estimates. However, both systems in the former case can be verified to be strongly detectable, thus, resilient estimates can be obtained in this case with less vulnerable sensors, as one may expect.

7.4 Attack Detection and Identification

Next, we address Problem 1.3 by investigating how the properties of the resilient state estimation algorithm in Sect. 7.3.2 affect attack detection and identification.

To begin, it is worth recalling that the resilient state estimation algorithms in the previous section are indifferent about whether the switching and false data injection attacks on the system are strategic. Nonetheless, it is critical to understand how our algorithms can detect or identify strategic attacks. In particular, we consider strategic attackers who aim to deceive the system operator/defender into believing that the mode of operation is $q \in \mathcal{Q}$, $q \neq q^*$, by means of selecting data injection signals d_k and the true mode $q^* \in \mathscr{Q}$. We call an attack *unidentifiable*, if the system operathor is not able to reconstruct/identify it. Moreover, the attack is *undetectable*, if it is *unidentifiable* and is unnoticeable. Below, we formally define the concepts of attack detection and attack identification, which are extensions of their counterparts in Yong et al. (2018) [Definitions 5.1 & 5.2].

Definition 7.2 (Switching and Data Injection Attack Detection) A switching and data injection attack is detected if the true mode $q^* \in \mathscr{Q}$ (chosen by attacker) has the maximum mean probability when using the resilient state estimation algorithm in Algorithm 7.2 or is not distinguishable from another mode $q \in \mathscr{Q}, q \neq q^*$ (chosen by defender) on average, in the presence of the stochastic/aleatoric uncertainty (i.e., if Assumption 7.1 holds), or if it is not eliminated by applying Algorithm 7.4 in the presence of the set-valued/epistemic uncertainty (i.e., if Assumption 7.2 holds).

Definition 7.3 (Switching and Data Injection Attack Identification) A switching and data injection attack strategy is identified if the attack is detected and in addition, the true mode $q^* \in \mathscr{Q}$ is uniquely determined on average (under aleatoric/stochastic uncertainty) or all false modes are eliminated (under epistemic/set-valued uncertainty), which reveals that the *mode attack* and *signal attack location*, and asymptotically unbiased estimates and/or uniformly bounded set-valued estimates of attack signals d_k can be obtained, i.e., the *signal magnitude attack* is reliably estimated.

It is obvious from the definitions above that if an attack is undetectable, it is also unidentifiable. Equivalently, if an attack is identifiable, then it is detectable. It is worth noting however that attack detection or identification is not required for calculating resilient state estimates. For example, in the simple case where there are no attacks, i.e., $d_k = 0$ for all k, the performance of state estimates of all models will be equally good, meaning that the attacks need not be detected or identified in order to obtain resilient state estimates.

7.4.1 Attack Detection

Our resilient state estimation approach (i.e., Algorithms 7.2 and 7.4) guarantees that an attack will always be detected by Definition 7.2 for all three uncertainty models. This is formally stated through the following theorem, which is a generalization of [Yong et al. (2018), Theorem 5.3].

Theorem 7.3 (Attack Detection) *The resilient state estimation algorithms in Algorithms 7.2 (with ratios of prior being identically 1) and 4 guarantee that switching and data injection attacks are always detectable, for all three uncertainty models.*

Proof First, note that if Assumption 7.2 holds, i.e., in the presence of distribution-free and norm-bounded noise signals, by (7.9), (7.10) and Proposition 7.1, $\|r_k^{q*}\|_2 \leq \delta_{r,k}^{q,*} \leq \hat{\delta}_{r,k}^{q*}$, i.e., (7.11) never holds for $q = q^*$ and hence, q^* is never eliminated. On the other hand, if Assumption 7.1 holds, i.e., in the presence of Gaussian noise signals, since the Kullback Leibler divergence $D(f_\ell^* \| f_\ell^q)$ is greater than or equal to zero with equality if and only if $f_\ell^* = f_\ell^q$ ([Kullback and Leibler (1951), Lemma 3.1]), with $j = q^* \in \mathscr{Q}$ as the true model and $i \in \mathscr{Q}, i \neq q^*$, the summand in the exponent of the ratio of geometric means whose expression is given in Yong et al. (2021)[Lemma 14] is always non-negative, i.e., $D(f_\ell^* \| f_\ell^i) - D(f_\ell^* \| f_\ell^*) = D(f_\ell^* \| f_\ell^i) \geq 0$. In other words, the ratio of the true model mean probability to the model mean probabilities of any other mode ($i \in \mathscr{Q}, i \neq q^*$) cannot decrease and can at best remain the same as the ratio of their priors being one by assumption. Thus, either the true model is identified or both modes are indistinguishable and a flag can be raised for attack detection. ∎

7.4.2 *Attack Identification*

A combination of switching and false data injection attacks may not be identifiable, even if it is detectable. On the other hand, it directly follows from Definition 7.3 that the mode detectability/mean consistency is sufficient to identify an attack strategy/action. This is formalized via the following theorem.

Theorem 7.4 (Attack Identification) *Suppose mode detectability and/or mean consistency, i.e., Yong et al. (2021), Theorem 8 and/or Khajenejad and Yong (2019), Theorem 4 hold (and hence Yong et al. 2021, Corollary 13 also holds). Then, the switching and data injection attack strategy can be identified using the resilient state estimation algorithms in Algorithms 7.1–7.4.*

7.4.2.1 Sufficient or Necessary Condition for Unidentifiable Attacks

Under the stochastic uncertainty model (cf. Assumption 7.1), if the true mode is in the set of models and even if the estimator is not mean consistent, a sufficient condition for an attack signal to be unidentifiable was derived in our previous work (Yong et al. (2018)), which we recap here for the sake of completeness (for more details, see [Yong et al. (2018), Sect. 5.2]).

Theorem 7.5 (Unidentifiable Attack) *[Yong et al. (2018), Theorem 5.5] If Assumption 7.1 or 7.3 hold, $\tilde{\Gamma}_k^q T_{a,2,k}^q H_k^*$ has linearly independent rows and there exists $q \neq q^* \in \mathscr{Q}$ such that*

$$\mathscr{D}_k^s \triangleq (\tilde{\Gamma}_k^q T_{a,2,k}^q H_k^*)^\dagger (S_k^* - \tilde{\Gamma}_k^q T_{a,2,k}^q (\mathbb{E}[\mu_k^{q|*} \mu_k^{q|*\top}] + R_k)(\tilde{\Gamma}_k^q T_{a,2,k}^q)^\top))(\tilde{\Gamma}_k^q T_{a,2,k}^q H_k^*)^{\dagger\top} \tag{7.16}$$

is positive definite ($\succeq 0$) for all k. Moreover, we assume that $\mu_0^* = \mu_0^q$. Then, the attack is unidentifiable if the attacker chooses this mode $q^* \neq q$ as well as the attack signal d_k as a Gaussian sequence

$$d_k \sim \mathscr{N}(d_k^d, \mathscr{D}_k^s), \quad \forall k \tag{7.17}$$

with $\mathscr{D}_k^s$ *defined in* (7.16) *and* d_k^d *is given by*

$$\begin{aligned} d_k^d \triangleq \mathbb{E}[d_k] &= -(\tilde{\Gamma}_k^q T_{a,2,k}^q H_k^*)^\dagger \tilde{\Gamma}_k^q T_{a,2,k}^q (C_k^{q^*} \mathbb{E}[x_k] - C_k^q \hat{x}_{a,k|k}^{\star,q} + (D_k^{q^*} - D_k^q)\mathbb{E}[u_k]) \\ &= -(\tilde{\Gamma}_k^q T_{a,2,k}^q H_k^*)^\dagger \tilde{\Gamma}_k^q T_{a,2,k}^q (C_k^{q^*} \hat{x}_{a,k|k}^{q^*} - C_k^q \hat{x}_{a,k|k}^{\star,q} + (D_k^{q^*} - D_k^q)\mathbb{E}[u_k]), \ \forall k. \end{aligned} \tag{7.18}$$

The above theorem highlights that an unidentifiable attack strategy often must rely on the existence of system "vulnerabilities" as well as the computational capability and system knowledge that are comparable to that of the system operator/defender. For the former factor, a system designer can consider these conditions as preventative mitigation guides for securing the system.

On the other hand, if Assumption 7.2 or 7.3 hold (i.e., epistemic/set-valued uncertainty is present), we provide a necessary condition for the attack signals to be unidentifiable, i.e., a condition that the attacker must ensure in order to guarantee that the attack signals are not identifiable.

Theorem 7.6 (A Necessary Condition for Unidentifiable Attacks) *Suppose Assumption 7.2 or 7.3 holds and* $T_{e,2,k}^q \neq T_{a,2,k}^{q'}, \forall k \geq 0, \forall q, q' \in \mathscr{Q}, q \neq q'$. *Then, a necessary condition for the attack signal to be unidentifiable is that it has limited energy when* $q = q^*$, *i.e.,* $\lim_{k\to\infty} \|d_{0:k}^{q*}\|_2 < \infty$, *where* $d_{0:k}^{q*} \triangleq \left[d_k^{q*\top} \ d_{k-1}^{q*\top} \ \ldots d_0^{q*\top}\right]^\top$.

Proof Using contraposition, suppose the attack signal has unlimited energy. Then, by [Khajenejad and Yong (2019), Theorem 4], all false modes will be eliminated after some large enough time step K and hence, the system is mode detectable (cf. Sect. 7.3.2.2). Thus, by Theorem 7.4, the attack strategy can be identified using the resilient state estimation algorithm and consequently, the attack signal cannot be unidentifiable. ■

This result has the important implication that attack signals must have limited energy to remain unidentifiable, and in this case, the harm that an attacker can inflict on a Cyber-Physical Systems (CPS) may also be limited. Note that the attack impact could still be catastrophic in this case, which incentives us to design attack mitigation approach in Sect. 7.5.

7.4.2.2 A Sufficient Condition for Resilient State Estimation

Finally, under the assumption of stochastic/aleatoric uncertainty (cf. Assumption 7.1), a sufficient condition can be found in Yong et al. (2018) to ensure that the state estimates are unbiased, even when the true mode is not uniquely determined and the attack signal cannot be estimated/identified, which is restated below.

Theorem 7.7 (Resilience Guarantee) *[Yong et al. (2018), Theorem 5.7] Suppose $H_k^q = H_k$ and $D_k^q = D_k$ for all $q \in \mathscr{Q}$. Moreover, for all $q, q' \in \mathscr{Q}$, if there exists T such that for all $k \geq T$ and the following hold*

(i) $\text{rank}\left[\tilde{\Gamma}_k^q T_{a,2,k}^q C_k^{q'} \ \tilde{\Gamma}_k^q T_{a,2,k}^q C_k^q\right] = 2\text{n}$, *if* $C_k^q \neq C_k^{q'}$,

(ii) $\text{rank}(\tilde{\Gamma}_k^q T_{a,2,k}^q C_k^{q'}) = \text{rank}(\tilde{\Gamma}_k^q T_{a,2,k}^q C_k^q) = n$, *if* $C_k^q = C_k^{q'}$,

then the state estimates obtained using Algorithm 7.2 are guaranteed to be resilient (i.e., asymptotically unbiased).

7.5 Attack Mitigation

We now move on to the challenge of minimizing the impact of attacks, i.e., attack mitigation (Problem 1.4), which is a step beyond attack detection and identification. In particular, we investigate the problem of rejecting/canceling data injection attacks assuming that the attack mode can be detected (thus, the superscript q is omitted throughout this section), while using the resilient state estimates for $\mathscr{H}_\infty$ controller synthesis, in the sense of guaranteeing the boundedness of the expected/worst case states and minimizing the effect of the attack signals. To this end, we consider a linear *dynamic controller* with attack/disturbance rejection terms in the following form, where $\hat{x}_{k|k}$, $\hat{d}_{1,k}$, $\hat{d}_{2,k-1}$ are obtained from Algorithms 7.1 or 7.3:

$$\begin{array}{ll} x_{k+1}^c & = A_k^c x_k^c + B_k^c \tilde{y}_k, \\ u_k & = C_k^c x_k^c + D_k^c \tilde{y}_k, \end{array} \tag{7.19}$$

with $K_k^c \triangleq \begin{bmatrix} A_k^c & B_k^c \\ C_k^c & D_k^c \end{bmatrix}$ being the dynamic controller gain that will be designed, $\tilde{y}_k \triangleq \left[\hat{x}_{k|k}^\top \ \hat{d}_{1,k}^\top \ \hat{d}_{2,k-1}^\top\right]^\top$, $B_k^c \triangleq \left[B_{x,k}^c \ B_{d_1,k}^c \ B_{d_2,k}^c\right]$ and $D_k^c \triangleq \left[D_{x,k}^c \ D_{d_1,k}^c \ D_{d_2,k}^c\right]$. Note that we have used a delayed estimate of $d_{2,k-1}$ given in (7.2), which is the only estimate we can obtain in light of [Yong et al. (2016b), Eq. (6)]. Before designing K_k^c for the purpose of attack mitigation and stabilization, we first show that there exists a separation principle for linear discrete-time systems with unknown inputs (attacks), i.e., when the true mode is known, which allows us to design the controller gain K_k^c independently of the observer gain $\tilde{L}_k$.

Lemma 7.1 (Separation Principle) *The state feedback controller gain K_k^c in (7.19) can be designed independently of the state and input estimator gains $\tilde{L}_k$, $M_{1,k}$ and $M_{2,k}$ in Algorithms 7.1 and 7.3.*

Proof Using the dynamic controller (7.19) and the filter/observer equations in (7.2), (7.3) and (7.4), it can be verified that the system and controller states and the estimator error dynamics are given by

$$
\begin{aligned}
\begin{bmatrix} x^c_{k+1} \\ x_{k+1} \\ \tilde{x}_{k+1|k+1} \end{bmatrix} &= \begin{bmatrix} A^c_k & B^c_{x,k} & -B^c_{x,k} \\ B_k C^c_k & A_k + B_k D^c_{k,x} & -B D^c_{k,x} \\ 0 & 0 & (I - \tilde{L}_{k+1} C_{2,k})\overline{A}_k \end{bmatrix} \begin{bmatrix} x^c_k \\ x_k \\ \tilde{x}_{k|k} \end{bmatrix} \\
&+ \begin{bmatrix} B^c_{d_1,k} & B^c_{d_2,k} \\ G_{1,k} + B_k D^c_{d_1,k} & G_{2,k} + B_k D^c_{d_2,k} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} d_{1,k} \\ d_{2,k} \end{bmatrix} + \begin{bmatrix} -B^c_{d_1,k} & -B^c_{d_1,k} \\ -B_k D^c_{d_1,k} & -B_k D^c_{d_2,k} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} d_{1,k} - \hat{d}_{1,k} \\ d_{2,k} - \hat{d}_{2,k-1} \end{bmatrix} \qquad (7.20) \\
&+ \begin{bmatrix} 0 & 0 & 0 \\ I & 0 & 0 \\ (I - \tilde{L}_{k+1} C_{2,k+1}) & -(I - \tilde{L}_{k+1} C_{2,k+1}) & -(I - \tilde{L}_{k+1} C_{2,k+1}) \\ (I - G_{2,k} M_{2,k+1} C_{2,k+1}) & (I - G_{2,k} M_{2,k+1} C_{2,k+1}) G_{1,k} M_{1,k} & G_{2,k} M_{2,k+1} - \tilde{L}_{k+1} \end{bmatrix} \mathbf{w}_k,
\end{aligned}
$$

where $\mathbf{w}_k \triangleq \begin{bmatrix} w_k^\top \; v_{1,k}^\top \; v_{2,k+1}^\top \end{bmatrix}^\top$ and $\overline{A}_k \triangleq (I - G_{2,k-1} M_{2,k} C_{2,k})(A_k - G_{1,k} M_{1,k} C_{1,k})$. Since the state matrix has a block upper triangular structure, the eigenvalues of the controller and estimator are independent of each other, thus K^c_k and $\tilde{L}_k$ can be designed separately. ■

Armed with the above lemma, we present an $\mathscr{H}_\infty$ controller design for determining the controller gain matrix K^c_k that stabilizes the closed-loop system and mitigates the effects of attack signals.

Theorem 7.8 (Attack-Mitigating and Stabilizing $\mathscr{H}_\infty$Controller) *Suppose the system* (7.1) *is controllable in the true mode $q \in \mathscr{Q}$ (known or detected). Then, the dynamic controller in* (7.19) *mitigates the effects of data injection attacks and minimizes the $\mathscr{H}_\infty$-gain from the augmented noise signal $\tilde{w}_k$ to the state as the desired output, i.e., $\tilde{z}_k = x_k$, using feedback based on estimates $\tilde{y}_k \triangleq \begin{bmatrix} \hat{x}_k^\top \; \hat{d}_{1,k}^\top \; \hat{d}_{2,k-1}^\top \end{bmatrix}^\top$, where the gain matrix K^c_k is the $\mathscr{H}_\infty$-controller gain matrix that can be synthesized (e.g., using* `hinfsyn` *in MATLAB) for the following augmented system:*

$$
\begin{aligned}
\xi_{k+1} &= \tilde{A}_k \xi_k + \tilde{B}_{1,k} \tilde{w}_k + \tilde{B}_{2,k} u_k, \\
\tilde{z}_k &= \tilde{C}_{1,k} \xi_k + \tilde{D}_{11,k} \tilde{w}_k + \tilde{D}_{12,k} u_k, \\
\tilde{y}_k &= \tilde{C}_{2,k} \xi_k + \tilde{D}_{21,k} \tilde{w}_k + \tilde{D}_{22,k} u_k,
\end{aligned} \qquad (7.21)
$$

where $\tilde{A}_k \triangleq \begin{bmatrix} A_k & G_{1,k} & G_{2,k} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$, $\tilde{B}_{1,k} \triangleq \begin{bmatrix} I & 0 & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 & 0 \end{bmatrix}$, $\tilde{B}_{2,k} \triangleq \begin{bmatrix} B_k \\ 0 \\ 0 \end{bmatrix}$, $\tilde{C}_{1,k} \triangleq \begin{bmatrix} I & 0 & 0 \end{bmatrix}$, $\tilde{C}_{2,k} \triangleq \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}$, $\tilde{D}_{11,k} \triangleq \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$, $\tilde{D}_{12,k} \triangleq 0$, $\tilde{D}_{21,k} \triangleq \begin{bmatrix} 0 & 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & 0 & I \end{bmatrix}$ *and* $\tilde{D}_{22,k} \triangleq \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^\top$.

Proof By Lemma 7.1, the state feedback gain, K^c_k, can be independently designed with no effect on the stability of the resilient state estimator/observer. In other words, K^c_k can be chosen optimally, in the sense of an $\mathscr{H}_\infty$-controller such that the augmented closed-loop system is stable, and that the effects of the augmented noise $\tilde{w}_k$ on the desired controlled output $\tilde{z}_k \triangleq x_k$ are minimized. To achieve this, we consider the following augmented system:

$$\begin{aligned} x_{k+1} &= A_k x_k + B_k u_k + G_{1,k} d_{1,k} + G_{2,k} d_{2,k} + w_k, \\ d_{1,k+1} &= \tilde{w}_{1,k}, \\ d_{2,k+1} &= \tilde{w}_{2,k}, \end{aligned} \tag{7.22}$$

with the augmented state $\xi_k \triangleq \left[x_k^\top \; d_{1,k}^\top \; d_{2,k}^\top\right]^\top$, where the goal is to use the dynamic controller (7.19) with estimates/"observations" $\tilde{y}_k \triangleq \left[\hat{x}_k^\top \; \hat{d}_{1,k}^\top \; \hat{d}_{2,k-1}^\top\right]^\top$ to stabilize the desired output/state $\tilde{z}_k \triangleq x_k$, while minimizing the effect of the augmented noise signal $\tilde{w}_k \triangleq \left[w_k^\top \; \tilde{w}_{1,k}^\top \; \tilde{w}_{2,k}^\top \; \tilde{x}_{k|k} \; \tilde{d}_{1,k}^\top \; \tilde{d}_{2,k}^\top\right]^\top$. Then, by plugging the control input u_k from (7.19) into (7.22), we obtain (7.21), where an $\mathscr{H}_\infty$-controller can be synthesized to achieve the minimum $\mathscr{H}_\infty$ performance. It is worth re-emphasizing that the control synthesis process is completely independent of the observer gains $\tilde{L}_k$, $M_{1,k}$, $M_{2,k}$. ■

Remark 7.1 The dynamic feedback gain K_k^c can be synthesized using the command

$$[K_k^c, CL_k, \gamma_k] = \texttt{hinfsyn}(P, \text{size}(D_{22,k}, 1), \text{size}(D_{22,k}, 2))$$

in MATLAB, where $P \triangleq \begin{bmatrix} \tilde{A}_k & \tilde{B}_{1,k} & \tilde{B}_{2,k} \\ \tilde{C}_{1,k} & \tilde{D}_{11,k} & \tilde{D}_{12,k} \\ \tilde{C}_{2,k} & \tilde{D}_{21,k} & \tilde{D}_{22,k} \end{bmatrix}$.

7.6 Simulation Examples

7.6.1 Benchmark System (Signal Magnitude Location Attacks)

The resilient state estimation problem for a system (modified from Yong et al. (2016b) and has been used as a benchmark for several state and input filters/observers) is considered in this example, where there exists only one mode of operation ($t_m = 1$) as well as possible attacks on the actuator and four of the five sensors ($t_a = 1, t_s = 4$):

$$A=\begin{bmatrix}0.5&2&0&0&0\\0&0.2&1&0&1\\0&0&0.3&0&1\\0&0&0&0.7&1\\0&0&0&0&0.1\end{bmatrix};\quad B=G=\begin{bmatrix}1\\0.1\\0.1\\1\\0\end{bmatrix};\quad C=\begin{bmatrix}1&0&0&0&0\\0&1&-0.1&0&0\\0&0&1&-0.5&0.2\\0&0&0&1&0\\0&0.25&0&0&1\end{bmatrix};$$

$$H=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\\0&0&0&0\end{bmatrix};\quad Q=10^{-4}\begin{bmatrix}1&0&0&0&0\\0&1&0.5&0&0\\0&0.5&1&0&0\\0&0&0&1&0\\0&0&0&0&1\end{bmatrix};\quad R=10^{-4}\begin{bmatrix}1&0&0&0.5&0\\0&1&0&0&0.3\\0&0&1&0&0\\0.5&0&0&1&0\\0&0.3&0&0&1\end{bmatrix}.$$

We consider the known input $u_k = \begin{cases}2, & 100 \le k \le 300\\ -2, & 500 \le k \le 700\\ 0, & \text{otherwise}\end{cases}$, whereas the unknown inputs (attacks) are as depicted in Fig. 7.4. Moreover, we assume that there are at most $p = 4$ attacks with no constraints on n_a and n_s, and consequently, there are $\mathfrak{N} = 1 \cdot \binom{5}{4} = 5$ models. The signal attack locations alternate between $q = 3$ (attack on actuator and sensors 1, 3, 4) and $q = 2$ (attack on actuator and sensors 1, 2, 4) every 350s, i.e., the dwell time is 350s.

From the top plot in Fig. 7.3 that depicts the computed mode probabilities (under aleatoric Gaussian uncertainties), we observe that except during the short transients after $t = 350$s and $t = 700$s due to switching, the mode probabilities converge to their true values ($q^* = 3 \to q^* = 2 \to q^* = 3$). On the other hand, Fig. 7.3 (bottom) depicts the values of mode indicator index, $q \times i_q$ for each mode, over time, assuming epistemic bounded-norm distribution-free uncertainties, with i_q defined as

$$i_q \triangleq \begin{cases}0, & \text{if } q \text{ is eliminated},\\ 1, & \text{otherwise},\end{cases} \qquad \forall q \in \mathcal{Q}.$$

Hence, $q \times i_q$ equals q if the mode q is not eliminated and is zero otherwise. As expected, it can be observed from Fig. 7.3 (bottom) that except for $q = 3$ and $q = 2$, the other modes are eliminated after some time steps.

Figure 7.4 shows computed state and unknown attack *point* estimates for the case of aleatoric (stochastic) uncertainty model, as well *set-valued* sate and unknown input (attack) estimates, when epistemic (distribution-free and norm-bounded) uncertainty model is assumed. The point estimates are seen to be close to the true values, even before the mode probabilities converge, and both the point estimates and the actual values of the states and unknown inputs (attacks) are within the set estimates, which are uniformly bounded and convergent set sequences, as expected. Similar results (not shown for brevity) are obtained for all other attack modes, $q = 1$ (attack on actuator and sensors 1, 2, 3), $q = 4$ (attack on actuator and sensors 2, 3, 4) and $q = 5$ (attack on sensors 1, 2, 3, 4). Thus, this example illustrates that when switching

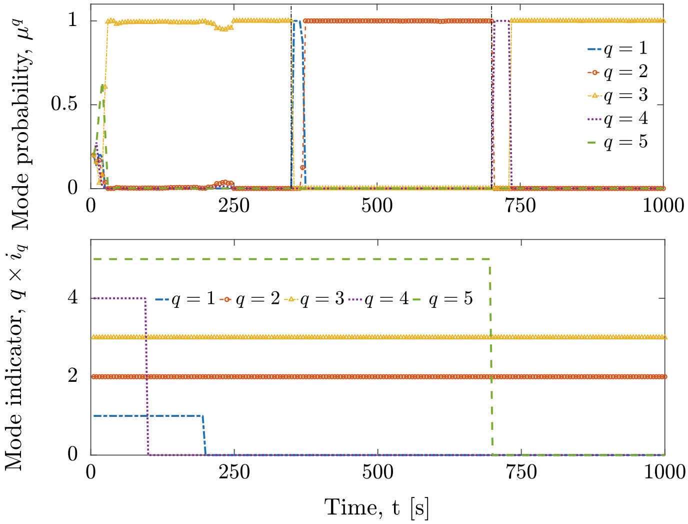

Fig. 7.3 Mode probabilities (top) assuming aleatoric/stochastic uncertainty model, as well as mode indicators (bottom) assuming epistemic/set-valued uncertainty model for the system in Sect. 7.6.1 with alternating switchings between $q = 3$ and $q = 2$ every 350s

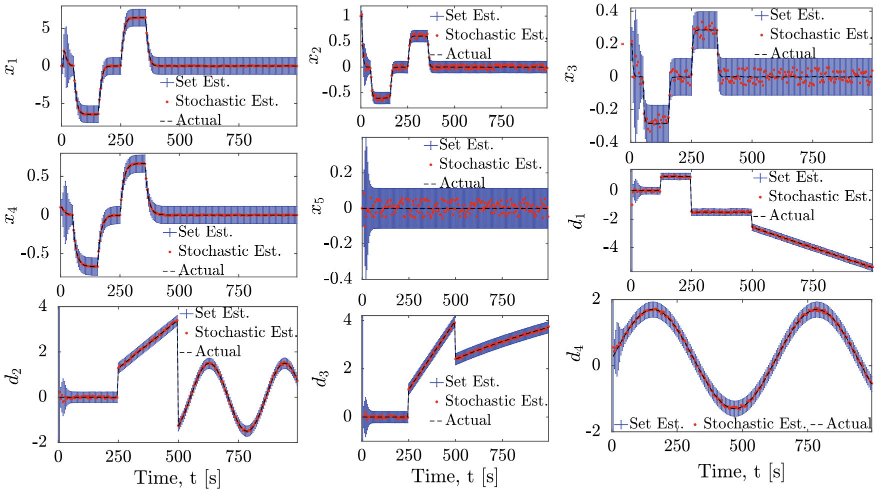

Fig. 7.4 State and attack magnitude estimates in Sect. 7.6.1 with switching between $q = 3$ and $q = 2$ with the dwell time 350 s

attacks and signal location attacks do not change quickly/frequently, i.e., the dwell time is large enough, our proposed methods work well.

7.6.2 IEEE 68-Bus Test System (Mode and Signal Magnitude Attacks)

The proposed algorithms are also applied to the IEEE 68-bus test system shown in [Yong et al. (2018), Fig. 7] to demonstrate their scalability to large systems, as well as to apply our attack mitigation approach.

An undirected graph $(\mathscr{V}, \mathscr{E})$ with the set of nodes (buses), $\mathscr{V} \triangleq \{1, \ldots, N\}$ and the set of edges (transmission/tie lines) $\mathscr{E} \subseteq \mathscr{V} \times \mathscr{V}$ is often used to describe a power network, where the busses may represent generator buses $i \in \mathscr{G}$, or load buses $i \in \mathscr{L}$. $\mathscr{S}_i \triangleq \{j \in \mathscr{V} \setminus \{i\} | (i, j) \in \mathscr{E}\}$ denotes the set of neighboring buses of $i \in \mathscr{V}$. In particular, there are 16 generator buses and 52 load buses for the IEEE 68-bus test system, i.e., $|\mathscr{G}| = 16$, $|\mathscr{L}| = 52$ and $|\mathscr{V}| = 68$. Similar to [Wood et al. (2013), Chap. 10], the dynamics of each bus, $i \in \mathscr{V}$, can be described by the following dynamical system:

$$\begin{aligned}\dot{\theta}_i(t) &= \omega_i(t),\\ \dot{\omega}_i(t) &= -\tfrac{1}{m_i}[D_i\omega_i(t) + \textstyle\sum_{j\in\mathscr{S}_i} P_{tie}^{ij}(t) - (P_{M_i}(t) + d_{a,i}(t)) + P_{L_i}(t) + w_i(t)],\end{aligned} \tag{7.23}$$

with the system states being the phase angle $\theta_i(t)$ and angular frequency $\omega_i(t)$ (hence, the state space dimension is $n = 136$) and an actuator attack signal $d_{a,i}(t)$. The power flow between neighboring buses i, j, such that $(i, j) \in \mathscr{E}$, is given by $P_{tie}^{ij}(t) = -P_{tie}^{ji}(t) = t_{ij}(\theta_i(t) - \theta_j(t))$, while $P_{M_i}(t)$ and $P_{L_i}(t)$ denote the mechanical power and power demand, respectively. The mechanical power $P_{M_i}(t)$ is the control input for the generator bus $i \in \mathscr{G}$ and is zero at load bus $i \in \mathscr{L}$. On the other hand, power demand $P_{L_i}(t)$ is taken as a known input since it can be calculated using load forecasting methods (e.g., Alfares and Nazeeruddin 2002). We assume that the noise $w_i(t)$ is a zero-mean truncated Gaussian signal (satisfying Assumption 7.3) with covariance matrix $Q_i(t) = 0.01$, $\eta_w = 0.03$ and the system parameters being adopted from Kundur et al. (1994) [p. 598]: $D_i = 1$, $t_{ij} = 1.5$ for all $i \in \mathscr{V}$, $j \in \mathscr{S}_i$ and $t_{ij} = 0$ otherwise. Angular momentums are $m_i = 10$ for $i \in \mathscr{G}$ and a larger value $m_i = 100$ for load buses $i \in \mathscr{L}$.

The measurements are sampled at discrete times (with sampling time $\Delta t = 0.01$s), satisfying the following output equation:

$$y_{i,k} = \left[P_{elec,i,k}\ \theta_{i,k}\ \omega_{i,k}\right]^\top + v_{i,k}, \tag{7.24}$$

where $P_{elec,i,k} = D_i\omega_{i,k} + P_{L_i,k}$ is the electrical power output and $v_{i,k}$ is a truncated zero-mean Gaussian noise signal with covariance matrix $R_i(t) = 0.01^4 I_3$ and $\eta_v =$

0.03. The continuous system dynamics (7.23) is also discretized with a sampling time of $\Delta t = 0.01$s. Furthermore, in this example, we choose the control inputs $P_{M_i,k}$ and $P_{L_i,k}$ through synthesizing an $\mathscr{H}_\infty$-optimal dynamic controller in the form of (7.19), as described in Theorem 7.8, to regulate the phase angles to $\theta_i = 10$ rad and mitigate the effect of the unknown attack signal.

As shown in Yong et al. (2018) [Fig. 7], the attacker could inject false data into the actuators and attack the transmission lines. Eight potential attack modes ($|\mathscr{Q}| = 8$) are considered:

Mode $q = 1$: Lines {27,53},{53,54},{60,61} & actuator $G1$.
Mode $q = 2$: Lines {18,49},{18,50} & actuator $G2$.
Mode $q = 3$: Line {40,41} & actuator $G3$.
Mode $q = 4$: Lines {18,49},{18,50},{27,53},{53,54},{60,61} & actuator $G4$.
Mode $q = 5$: Lines {27,53},{40,41},{53,54},{60,61} & actuator $G5$.
Mode $q = 6$: Lines {18,49},{18,50},{40,41} & actuator $G6$.
Mode $q = 7$: Lines {18,49},{18,50},{27,53},{40,41},{53,54},{60,61} & actuator $G7$.
Mode $q = 8$: Actuator $G8$.

We study a time-varying attack scenario where the attack mode is $q = 2$ for $t = [0, 2.5)$s followed by $q = 5$ for $t = [2.5, 5)$s, while the actuator attack signal is given in Fig. 7.6. Our goal is to demonstrate that attack signals can be detected, identified, and mitigated by our proposed approach. To synthesize the attack-mitigating dynamic controller in the form of (7.19), we consider three cases, depending on the three different assumptions on possible uncertainty models: (i) aleatoric/stochastic uncertainty (cf. Assumption 7.1), where we use $\hat{x}_{a,k|k}$ and $\hat{d}_{a,k-1}$ (i.e., the most likely estimates among all mode-matched estimates) returned by Algorithm 7.2 in (7.19), (ii) epistemic/bounded norm uncertainty (cf. Assumption 7.2), where we plug $\hat{x}_{e,k|k}$ and $\hat{d}_{e,k-1}$ (i.e., the centroids of the union of all the set-estimates that correspond to non-eliminated modes) returned by Algorithm 7.4 in (7.19), and (iii) combined uncertainty (cf. Assumption 7.3), where we use the most likely point (stochastic) estimates among all the ones that correspond to the non-eliminated modes, as described in Sect. 7.3.1.

Figure 7.5 demonstrates that attacks are detected almost instantaneously, and the attack modes are quickly identified. Further, Fig. 7.6 depicts the successful identification of the actuator attack signal and estimation of all system states (not depicted for brevity). Finally, the proposed attack mitigation is shown to be effective in regulating the phase angles at 10 rad/s despite attacks, while without attack mitigation, attackers can drastically influence the phase angles as shown in Fig. 7.6.

7.7 Conclusion

We addressed the problem of resilient state estimation for switching (mode/topology) attacks and attacks on actuator and sensor signals of Cyber-Physical Systems

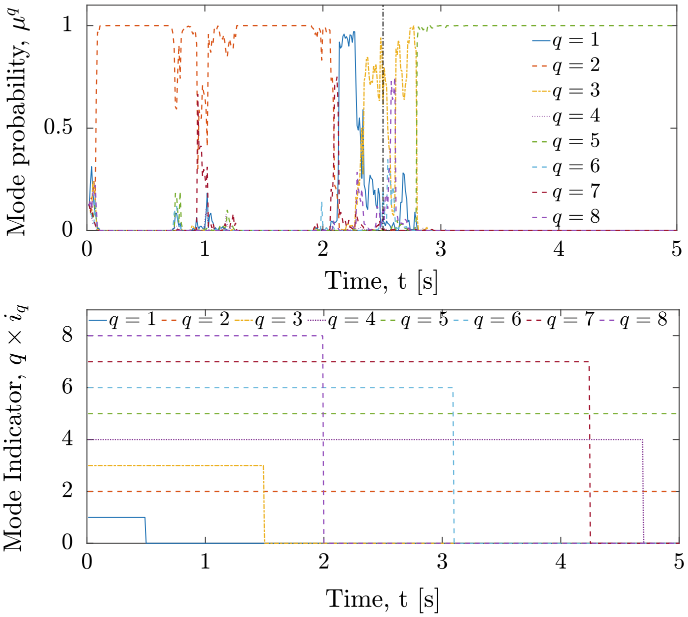

Fig. 7.5 Estimates of mode probabilities when the attack mode switches from $q = 2$ to $q = 5$ at 2.5s assuming stochastic uncertainties, as well as mode indicators assuming bounded norm uncertainties in Sect. 7.6.2

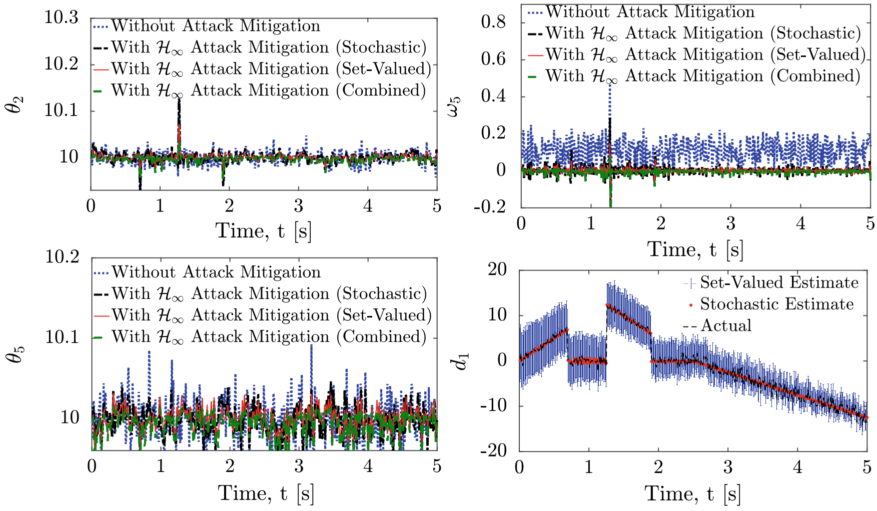

Fig. 7.6 A comparison of system states with and without the proposed attack mitigation, as well as the attack signal and its point-valued (stochastic) and set-valued (bounded-error) estimates

(CPS). We modeled the problem as a hidden-mode switched linear system with unknown inputs, where we considered three uncertainty models for the noise signals: (a) aleatoric/stochastic, (b) epistemic/set-valued and distribution-free, (c) truncated Gaussian uncertainties. We showed that the multiple-model inference algorithm in Yong et al. (2021); Khajenejad and Yong (2019) is a good solution to these problems. Furthermore, for the multiple-model approach, we presented an achievable upper bound on the maximum number of correctable signal attacks, as well as the maximum number of required models. We also derived sufficient conditions for attack (un-)detectability and identification and necessary conditions for the attack signal to be unidentifiable. Moreover, we designed an attack-mitigating $\mathscr{H}_\infty$-controller to minimize the effects of the attack signals. The effectiveness of our methods for resilient estimation, attack detection, and mitigation was demonstrated in simulations, including using an IEEE 68-bus test system.

Appendix

System Transformation

To obtain the mode-matched input and state estimator (7.2)–(7.4), we will consider a system transformation for the continuous system dynamics and output equation in (7.1) for each mode q (Yong et al. 2016b). First, we rewrite the direct feedthrough matrix H_k using singular value decomposition as $H_k = \begin{bmatrix} U_{1,k} & U_{2,k} \end{bmatrix} \begin{bmatrix} \Sigma_k & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_{1,k}^\top \\ V_{2,k}^\top \end{bmatrix}$, where $\Sigma_k \in \mathbb{R}^{p_{H_k} \times p_{H_k}}$ is a diagonal matrix of full rank, $U_{1,k} \in \mathbb{R}^{l \times p_{H_k}}$, $U_{2,k} \in \mathbb{R}^{l \times (l - p_{H_k})}$, $V_{1,k} \in \mathbb{R}^{p \times p_{H_k}}$ and $V_{2,k} \in \mathbb{R}^{p \times (p - p_{H_k})}$ with $p_{H_k} := \text{rk}(H_k)$, while $U_k := \begin{bmatrix} U_{1,k} & U_{2,k} \end{bmatrix}$ and $V_k := \begin{bmatrix} V_{1,k} & V_{2,k} \end{bmatrix}$ are unitary matrices. When there is no direct feedthrough, Σ_k, $U_{1,k}$ and $V_{1,k}$ are empty matrices,[3] and $U_{2,k}$ and $V_{2,k}$ are arbitrary unitary matrices.

Further, we define two orthogonal components of the unknown input d_k given by

$$d_{1,k} \triangleq V_{1,k}^\top d_k, \quad d_{2,k} \triangleq V_{2,k}^\top d_k. \tag{7.25}$$

Since V_k is unitary, $d_k = V_{1,k} d_{1,k} + V_{2,k} d_{2,k}$. Thus, the continuous system dynamics and output equation in (7.1) for each mode q can be rewritten as

$$x_{k+1} = A_k x_k + B_k u_k + G_{1,k} d_{1,k} + G_{2,k} d_{2,k} + w_k, \tag{7.26}$$

$$y_k = C_k x_k + D_k u_k + H_{1,k} d_{1,k} + v_k, \tag{7.27}$$

[3] We adopt the convention that the inverse of an empty matrix is also an empty matrix and assume that operations with empty matrices are possible.

where $G_{1,k} := G_k V_{1,k}$, $G_{2,k} := G_k V_{2,k}$, and $H_{1,k} := H_k V_{1,k} = U_{1,k}\Sigma_k$. Next, we decouple the output y_k using a nonsingular transformation $T_{a,k} = \begin{bmatrix} T_{a,1,k}^\top & T_{a,2,k}^\top \end{bmatrix}^\top = \begin{bmatrix} I_{p_{H_k}} & -U_{1,k}^\top R_k U_{2,k}(U_{2,k}^\top R_k U_{2,k})^{-1} \\ 0 & I_{(l-p_{H_k})} \end{bmatrix} \begin{bmatrix} U_{1,k}^\top \\ U_{2,k}^\top \end{bmatrix}$ in the presence of aleatoric uncertainty, i.e., if Assumption 7.1 holds, $T_{e,k} = \begin{bmatrix} T_{e,1,k}^\top & T_{e,2,k}^\top \end{bmatrix}^\top = \begin{bmatrix} U_{1,k} & U_{2,k} \end{bmatrix}^\top$ in the presence of epistemic uncertainty, i.e., if Assumption 7.2 holds, and both in the presence of truncated Gaussian uncertainty, i.e., if Assumption 7.3 holds. Consequently, we obtain $z_{t,1,k} \in \mathbb{R}^{p_{H_k}}$ and $z_{t,2,k} \in \mathbb{R}^{l-p_{H_k}}$, $\forall t \in \{a, e\}$, as

$$\begin{aligned} z_{t,1,k} &\triangleq T_{t,1,k} y_k = C_{t,1,k} x_k + D_{t,1,k} u_k + \Sigma_k d_{1,k} + v_{t,1,k}, \\ z_{t,2,k} &\triangleq T_{t,2,k} y_k = C_{t,2,k} x_k + D_{t,2,k} u_k + v_{t,2,k}, \end{aligned} \tag{7.28}$$

where $C_{t,1,k} \triangleq T_{t,1,k} C_k$, $C_{t,2,k} \triangleq T_{t,2,k} C_k = U_{2,k}^\top C_k$, $D_{t,1,k} \triangleq T_{t,1,k} D_k$, $D_{t,2,k} \triangleq T_{t,2,k} D_k = U_{2,k}^\top D_k$, $v_{t,1,k} \triangleq T_{t,1,k} v_k$, and $v_{t,2,k} \triangleq T_{t,2,k} v_k = U_{2,k}^\top v_k$. This system transformation essentially decouples the output equation involving y_k into two components, one with a full rank direct feedthrough matrix and the other without direct feedthrough. The transformation is also chosen such that in the case of aleatoric uncertainty, the measurement noise terms for the decoupled outputs are uncorrelated. The covariances of $v_{1,k}$ and $v_{2,k}$ are

$$\begin{aligned} R_{1,k} &\triangleq \mathbb{E}[v_{1,k} v_{1,k}^\top] = T_{a,1,k} R_k T_{a,1,k}^\top \succ 0, \\ R_{2,k} &\triangleq \mathbb{E}[v_{2,k} v_{2,k}^\top] = T_{a,2,k} R_k T_{a,2,k}^\top = U_{2,k}^\top R_k U_{2,k} \succ 0, \\ R_{12,k} &\triangleq \mathbb{E}[v_{1,k} v_{2,k}^\top] = T_{a,1,k} R_k T_{a,2,k}^\top = 0, \\ R_{12,(k,i)} &\triangleq \mathbb{E}[v_{1,k} v_{2,i}^\top] = T_{a,1,k} \mathbb{E}[v_k v_i^\top] T_{a,2,i}^\top = 0, \ \forall k \neq i. \end{aligned} \tag{7.29}$$

Moreover, $v_{1,k}$ and $v_{2,k}$ are uncorrelated with the initial state x_0 and process noise w_k. Further, in the case of bounded-norm uncertainty, the transform is also chosen such that $\| \begin{bmatrix} v_{1,k}^\top & v_{2,k}^\top \end{bmatrix}^\top \| = \| \begin{bmatrix} U_{1,k} & U_{2,k} \end{bmatrix}^\top v_k \| = \|v_k\|$.

Residual Upper Bounds

The upper bounds on the residual signal in Proposition 7.2 can be found as in Khajenejad and Yong (2019) [Theorem 3]:

$$\begin{aligned} \delta_{r,k}^{q,inf} &\triangleq \|\mathbb{A}_k^q t_k^\star\|_2, \\ \delta_{r,k}^{q,tri} &\triangleq \delta_0^{x,q} \|C_{e,2,k}^q \overline{A}_k^q {A_{e,k}^q}^{k-1}\|_2 + \eta_w (\|C_{e,2,k}^q \overline{A}_k^q {A_{e,k}^q}^{k-2}\|_2 + \|C_{e,2,k}^q B_{e,w,k}^{\star,q}\|_2) \\ &+ \textstyle\sum_{i=1}^{k-2} [\eta_w \|C_{e,2,i}^q \overline{A}_i^q {A_{e,i}^q}^{i} B_{e,w,i}^q\|_2 + \eta_v \|C_{e,2,i}^q \overline{A}_i^q {A_{e,i}^q}^{i} (B_{e,v_1,i}^q + A_{e,i}^q B_{e,v_2,i}^q)\|_2] \\ &+ \eta_v (\|C_{e,2,k}^q \overline{A}_k^q {A_{e,k}^q}^{k-2} B_{e,v_1,k}^q\|_2 + \|C_{e,2,k}^q (B_{e,v_1,k}^{q,\star} + \overline{A}_k^q B_{e,v_2,k}^q)\|_2 \\ &+ \|C_{e,2,k}^q B_{e,v_2,k}^{q,\star} + T_{e,2,k}^q\|_2), \end{aligned} \tag{7.30}$$

where $t_k^\star$ is a vertex of the following hypercube:

$$\mathscr{X}_k^q \triangleq \left\{x \in \mathbb{R}^{(n+l)(k+1)} \ : \ |x(i)| \le \begin{cases} \delta_0^x, 1 \le i \le n \\ \eta_w, n+1 \le i \le n(k+1) \\ \eta_v, n(k+1)+1 \le i \le (n+l)(k+1) \end{cases} \right\},$$

i.e.,

$$t_k^\star(i) \in \begin{cases} \{-\delta_0^x, \delta_0^x\}, 1 \le i \le n, \\ \{-\eta_w, \eta_w\}, n+1 \le i \le n(k+1), \\ \{-\eta_v, \eta_v\}, n(k+1)+1 \le i \le (n+l)(k+1) \end{cases} \qquad \text{and}$$

$$\overline{A}_k \triangleq \Phi_k \hat{A}_k, \ V_{e,k} \triangleq V_{1,k} M_{1,k} C_{1,k} + V_{2,k} M_{2,k} C_{2,k} \hat{A}_k, \ A_{e,k} \triangleq (I - \tilde{L}_k C_{2,k}) \overline{A}_k,$$
$$B_{e,w,k} \triangleq (I - \tilde{L}_k C_{2,k}) \Phi_k, \ B_{e,v_1,k} \triangleq -(I - \tilde{L}_k C_{2,k}) \Phi_k G_{1,k} M_{1,k} T_{1,k},$$
$$B_{e,v_2,k} \triangleq -((I - \tilde{L}_k C_{2,k}) G_{2,k} M_{2,k} + \tilde{L}_k) T_{2,k}.$$

References

H. Alfares, M. Nazeeruddin, Electric load forecasting: literature survey and classification of methods. Int. J. Syst. Sci. **33**(1), 23–34 (2002)

Y. Bar-Shalom, X. Li, T. Kirubarajan, *Estimation with Applications to Tracking and Navigation: Theory Algorithms and Software* (Wiley, 2004)

Y. Bar-Shalom, X. Li, T. Kirubarajan, *Estimation with Applications to Tracking and Navigation: Theory Algorithms and Software* (Wiley, 2004)

F. Blanchini, M. Sznaier, A convex optimization approach to synthesizing bounded complexity ℓ^∞ filters. IEEE Trans. Autom. Control **57**(1), 216–221 (2012)

H. Bodlaender, P. Gritzmann, V. Klee, J. Van Leeuwen, Computational complexity of norm-maximization. Combinatorica **10**(2), 203–225 (1990)

A.Cárdenas, S. Amin, S. Sastry, Research challenges for the security of control systems, in *Proceedings of the 3rd Conference on Hot Topics in Security*, ser. HOTSEC'08 (2008), pp. 6:1–6:6

A.Cárdenas, S. Amin, S. Sastry, Secure control: towards survivable cyber-physical systems, in *International Conference on Distributed Computing Systems Workshops* (2008), pp. 495–500

L. Cómbita, J. Giraldo, A. Cárdenas, N. Quijano, Response and reconfiguration of cyber-physical control systems: a survey, in *IEEE Colombian Conference on Automatic Control (CCAC)* (2015), pp. 1–6

M. Dahleh, I. Diaz-Bobillo, *Control of Uncertain Systems: a Linear Programming Approach* (Prentice-Hall, Inc., 1994)

G. Dan, H. Sandberg, Stealth attacks and protection schemes for state estimators in power systems, in *IEEE International Conference on Smart Grid Communications (SmartGridComm)* (2010), pp. 214–219

G. De Nicolao, G. Sparacino, C. Cobelli, Nonparametric input estimation in physiological systems: problems, methods, and case studies. Automatica **33**(5), 851–870 (1997)

J. Farwell, R. Rohozinski, Stuxnet and the future of cyber war. Survival **53**(1), 23–40 (2011)

H. Fawzi, P. Tabuada, S. Diggavi, Secure estimation and control for cyber-physical systems under adversarial attacks. IEEE Trans. Autom. Control **59**(6), 1454–1467 (2014)

B. Ghena, W. Beyer, A. Hillaker, J. Pevarnek, J. Halderman, Green lights forever: Analyzing the security of traffic infrastructure, in *8th USENIX Workshop on Offensive Technologies*, vol. 14, pp. 7–7 (2014)

S. Gillijns, B. De Moor, Unbiased minimum-variance input and state estimation for linear discrete-time systems. Automatica **43**(1), 111–116 (2007)

S. Gillijns, B. De Moor, Unbiased minimum-variance input and state estimation for linear discrete-time systems with direct feedthrough. Automatica **43**(5), 934–937 (2007)

R. Goebel, R. Sanfelice, A. Teel, Hybrid dynamical systems. IEEE Control Syst. Mag. **29**(2), 28–93 (2009)

X. Jin, W. Haddad, An adaptive control architecture for leader-follower multi-agent systems with stochastic disturbances and sensor and actuator attacks. Int. J. Control **92**(11), 2561–2570 (2019)

X. Jin, W. Haddad, Adaptive control for multi-agent systems with sensor-actuator attacks and stochastic disturbances. J. Guid. Control Dyn. **43**(1), 15–29 (2020)

X. Jin, W. Haddad, T. Yucelen, An adaptive control architecture for mitigating sensor and actuator attacks in cyber-physical systems. IEEE Trans. Autom. Control **62**(11), 6058–6064 (2017)

M. Khajenejad, S.Z. Yong, Simultaneous mode, input and state set-valued observers with applications to resilient estimation against sparse attacks, in *2019 IEEE 58th Conference on Decision and Control (CDC)* (IEEE, 2019), pp. 1544–1550

J. Kim, L. Tong, On topology attack of a smart grid: undetectable attacks and countermeasures. IEEE J. Sel. Areas Commun. **31**(7), 1294–1305 (2013)

P. Kitanidis, Unbiased minimum-variance linear state estimation. Automatica **23**(6), 775–778 (1987)

O. Kosut, L. Jia, R. Thomas, L. Tong, Malicious data attacks on the smart grid. IEEE Trans. Smart Grid **2**(4), 645–658 (2011)

S. Kullback, R. Leibler, On information and sufficiency. Ann. Math. Stat. **22**, 49–86 (1951)

P. Kundur, N.J. Balu, M.G. Lauby, *Power System Stability and Control* (McGraw-Hill New York, 1994)

C. Kwon, W. Liu, I. Hwang, Security analysis for cyber-physical systems against stealthy deception attacks, in *IEEE American Control Conference (ACC)* (2013), pp. 3344–3349

D. Li, S. Martínez, High-confidence attack detection via Wasserstein-metric computations. IEEE Control Syst. Lett. **5**(2), 379–384 (2020)

G. Liang, J. Zhao, F. Luo, S. Weller, Z.Y. Dong, A review of false data injection attacks against modern power systems. IEEE Trans. Smart Grid **8**(4), 1630–1638 (2017)

Y. Liu, P. Ning, M. Reiter, False data injection attacks against state estimation in electric power grids. ACM Trans. Inf. Syst. Secur. (TISSEC) **14**(1), 13 (2011)

S. Liu, S. Mashayekh, D. Kundur, T. Zourntos, K. Butler-Purry, A framework for modeling cyber-physical switching attacks in smart grid. IEEE Trans. Emerging Top. Comput. **1**(2), 273–285 (2013)

C. Ma, D. Yau, X. Lou, N. Rao, Markov game analysis for attack-defense of power networks under possible misinformation. IEEE Trans. Power Syst. **28**(2), 1676–1686 (2013)

E. Mazor, A. Averbuch, Y. Bar-Shalom, J. Dayan, Interacting multiple model methods in target tracking: a survey. IEEE Trans. Aerosp. Electron. Syst. **34**(1), 103–123 (1998)

J. Milošević, D. Umsonst, H. Sandberg, K. Johansson, Quantifying the impact of cyber-attack strategies for control systems equipped with an anomaly detector, in *European Control Conference (ECC)* (IEEE, 2018), pp. 331–337

S. Mishra, Y. Shoukry, N. Karamchandani, S. Diggavi, P. Tabuada, Secure state estimation: optimal guarantees against sensor attacks in the presence of noise, in *IEEE International Symposium on Information Theory (ISIT)* (2015), pp. 2929–2933

Y. Mo, B. Sinopoli, False data injection attacks in control systems, in *First Workshop on Secure Control Systems, CPS Week* (2010)

C. Murguia, J. Ruths, Cusum and chi-squared attack detection of compromised sensors, in *2016 IEEE Conference on Control Applications (CCA)* (IEEE, 2016), pp. 474–480

Y. Nakahira, Y. Mo, Attack-resilient $\mathscr{H}_2$, $\mathscr{H}_\infty$, and ℓ_1 state estimator (2018), arXiv:1803.07053

M. Pajic, J. Weimer, N. Bezzo, P. Tabuada, O. Sokolsky, I. Lee, G. Pappas, Robustness of attack-resilient state estimators, in *ACM/IEEE Intl. Conference on Cyber-Physical Systems* (2014), pp. 163–174

M. Pajic, P. Tabuada, I. Lee, and G. Pappas, Attack-resilient state estimation in the presence of noise, in *IEEE Conference on Decision and Control* (2015), pp. 5827–5832

F. Pasqualetti, F. Dörfler, F. Bullo, Attack detection and identification in cyber-physical systems. IEEE Trans. Autom. Control **58**(11), 2715–2729 (2013)

R. Patton, R. Clark, P. Frank, *Fault Diagnosis in Dynamic Systems: Theory and Applications*, ser. Prentice-Hall International Series in Systems and Control Engineering (Prentice Hall, 1989)

V. Renganathan, N. Hashemi, J. Ruths, T. Summers, Higher-order moment-based anomaly detection. IEEE Control Syst. Lett. (2021)

G. Richards, Hackers versus slackers. Eng. Technol. **3**(19), 40–43 (2008)

J. Shamma, K. Tu, Set-valued observers and optimal disturbance rejection. IEEE Trans. Autom. Control **44**(2), 253–264 (1999)

J. Slay, M. Miller, Lessons learned from the Maroochy water breach, in *International Conference on Critical Infrastructure Protection* (Springer, 2007), pp. 73–82

S. Sundaram, C. Hadjicostis, Delayed observers for linear systems with unknown inputs. IEEE Trans. Autom. Control **52**(2), 334–339 (2007)

J. Weimer, S. Kar, K. Johansson, Distributed detection and isolation of topology attacks in power networks, in *Proceedings of the 1st International Conference on High Confidence Networked Systems*, ser. HiCoNS '12 (ACM, 2012), pp. 65–72

A. Wood, B. Wollenberg, G. Sheble, *Power Generation, Operation, and Control* (Wiley, 2013)

M. Yadegar, N. Meskin, W. Haddad, An output-feedback adaptive control architecture for mitigating actuator attacks in cyber-physical systems. Int. J. Adapt. Control Signal Process. **33**(6), 943–955 (2019)

S.Z. Yong, Simultaneous input and state set-valued observers with applications to attack-resilient estimation, in *Annual American Control Conference (ACC)*. (IEEE, 2018), pp. 5167–5174

S.Z. Yong, M. Zhu, E. Frazzoli, Resilient state estimation against switching attacks on stochastic cyber-physical systems, in *Proceedings of the American Control Conference*, submitted (2015)

S.Z. Yong, M. Foo, E. Frazzoli, Robust and resilient estimation for cyber-physical systems under adversarial attacks, in *IEEE American Control Conference (ACC)* (2016a), pp. 308–315

S.Z. Yong, M. Zhu, E. Frazzoli, A unified filter for simultaneous input and state estimation of linear discrete-time stochastic systems. *Automatica*, vol. 63 (2016b), pp. 321–329. Extended version first appeared in September 2013 and is available from: http://arxiv.org/abs/1309.6627

S.Z. Yong, M. Zhu, E. Frazzoli, Simultaneous input and state estimation for linear time-varying continuous-time stochastic systems. IEEE Trans. Autom. Control **62**(5), 2531–2538 (2017)

S.Z. Yong, M. Zhu, E. Frazzoli, Switching and data injection attacks on stochastic cyber-physical systems: modeling, resilient estimation, and attack mitigation. ACM Trans. Cyber-Phys. Syst. **2**(2), 1–2 (2018)

S.Z. Yong, M. Zhu, E. Frazzoli, Simultaneous mode, input and state estimation for switched linear stochastic systems. Int. J. Robust Nonlinear Control **31**(2), 640–661 (2021)

K. Zetter, *Inside the Cunning, Unprecedented Hack of Ukraine's Power Grid* (Wired Magazine, 2016)

M. Zhu, S. Martínez, Stackelberg-game analysis of correlated attacks in cyber-physical systems, in *IEEE American Control Conference (ACC)* (2011), pp. 4063–4068

M. Zhu, S. Martínez, On distributed constrained formation control in operator-vehicle adversarial networks. Automatica **49**(12), 3571–3582 (2013)

Q. Zhu, T. Basar, Game-theoretic methods for robustness, security, and resilience of cyberphysical control systems: games-in-games principle for optimal cross-layer resilient control systems. IEEE Control Syst. **35**(1), 46–65 (2015)