University Library



A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Yang, T;Murguia, C;Kuijper, M;Nesic, D

Title:
A Multi-Observer Approach for Attack Detection and Isolation of Discrete-Time Nonlinear Systems

Date:
2018

Citation:
Yang, T., Murguia, C., Kuijper, M. & Nesic, D. (2018). A Multi-Observer Approach for Attack Detection and Isolation of Discrete-Time Nonlinear Systems. Proceedings of the 2018 Australian & New Zealand Control Conference, 00, pp.346-351. IEEE. https://doi.org/10.1109/ANZCC.2018.8606587.

Persistent Link:
https://hdl.handle.net/11343/251780

A Multi-Observer Approach for Attack Detection and Isolation of Discrete-Time Nonlinear Systems

Tianci Yang, Carlos Murguia, Margreta Kuijper, Dragan Nešić

***Abstract*— We address the problem of attack detection and isolation for a class of discrete-time nonlinear systems under (potentially unbounded) sensor attacks and measurement noise. We consider the case when a subset of sensors is subject to additive false data injection attacks. Using a bank of observers, each observer leading to an Input-to-State Stable (ISS) estimation error, we propose two algorithms for detecting and isolating sensor attacks. These algorithms make use of the ISS property of the observers to check whether the trajectories of observers are "consistent" with the attack-free trajectories of the system. Simulations results are presented to illustrate the performance of the proposed algorithms.**

I. Introduction

Traditional control systems composed of interconnected controllers, sensors, and actuators use point-to-point communication architectures. This is no longer suitable when new requirements – such as modularity, decentralisation of control, integrated diagnostics, quick and easy maintenance, and low cost – are necessary. To meet these requirements, Networked Control Systems (NCSs) have emerged as a technology that combines control, communication, and computation, and offers the necessary flexibility to meet new demands in distributed and large scale systems.

Recently, security of NCSs has become a very important issue as wireless communication networks might serve as new access points for adversaries trying to disrupt the system dynamics. Cyber-physical attacks on control systems have caused substantial damage to a number of physical processes. One of the most well-known examples is the attack on Maroochy Shire Council's sewage control system in Queensland, Australia that happened in January 2000. The attacker hacked into the controllers that activate and deactivate valves and caused flooding of the grounds of a hotel, a park, and a river with a million liters of sewage. Another incident is the very recent SuxNet virus that targeted Siemens' supervisory control and data acquisition systems which are used in many industrial processes. These incidents show that strategic mechanisms to identify and deal with attacks on NCSs are strongly needed.

In [1]-[22], a range of topics related to security of control systems have been discussed. In general, they provide analysis tools for quantifying the performance degradation induced by different classes of attacks and propose reaction strategies to counter their effect on the system dynamics. Most of the existing work, however, has considered control systems with linear dynamics, although in most engineering applications the dynamics of the plants being monitored and controlled is highly nonlinear. There are some results addressing the nonlinear case though. In [23], exploiting sensor redundancy, the authors address the problem of sensor attack detection and state estimation for uniformly observable continuous-time nonlinear systems. Similarly, in [24], the problem of state estimation and attack isolation for a class of noisy discrete-time nonlinear system is considered. In particular, the authors propose an observer-based estimator, using a bank of circle-criterion observers, which provides a robust estimate of the system state in spite of sensor attacks and measurement noise, and an estimator-based isolation algorithm without knowing the noise bounds. In this manuscript, we address the problem of attack detection and isolation of a class of discrete-time nonlinear systems in the presence of measurement noise and sensor attacks. We assume that bounds on the measurement noise and an upper bound on the number of attacked sensors are known. We consider the setting when the system has p sensors, all of which are subject to measurement noise and up to $q < p$ of them are attacked. We assume that q is known but the exact subset of sensors being attacked is unknown. Using a bank of the observers, each observer leading to an ISS estimation error, we propose two algorithms for detecting and isolating false data sensor attacks. These algorithms make use of the ISS property of the observers to check whether the trajectories of observers are "consistent" with the attack-free trajectories of the system. The main idea behind our algorithms is the following. Each observer in the bank is driven by a different subset of sensors. Thus, without attacks, the observers produce ISS estimation errors with respect to measurement noise only. For every pair of observers in the bank, we compute the largest difference between their estimates. If a pair of observers is driven by a subset of attack-free sensors, then the largest difference between their estimates is also ISS with respect to measurement noise only. However, if there are attacks on some of the sensors, the observers driven by those sensors might produce larger differences than the attack-free ones. These ideas work well under the assumption that less than $p/2$ sensors are attacked, i.e, $q < p/2$.

Notation.

We denote the set of real numbers by $\mathbb{R}$, the set of natural numbers by $\mathbb{N}$, the set of integers by $\mathbb{Z}$, and $\mathbb{R}^{n\times m}$ the set of $n \times m$ matrices for any $m, n \in \mathbb{N}$.

This work was supported by the Australian Research Council under the Discovery Project DP170104099.

The authors are with the Department of Electrical and Electronics Engineering, the University of Melbourne, Australia. tianciy@student.unimelb.edu.au

For any vector $v \in \mathbb{R}^n$, v_J denotes the stacking of all v_i, $i \in J$ and $J \subset \{1,\dots,n\}$, $|v| = \sqrt{v^\top v}$ and $\mathrm{supp}(v) = \{i \in \{1,\dots,n\} \,|\, v_i \neq 0\}$. For a sequence of vectors $\{v(k)\}_{k=0}^{\infty}$, we denote by $v_{[0,k]}$ the sequence of vectors $v(i)$, $i = 0,\dots,k$, $||v||_\infty \triangleq \sup_{k\geq 0} |v(k)|$ and $||v||_T \triangleq \sup_{0\leq k\leq T} |v(k)|$. We say a sequence $\{v(k)\} \in l_\infty$ if $||v||_\infty < \infty$. The binomial coefficient is denoted as $\binom{a}{b}$, where a, b are nonnegative integers. We denote the cardinality of a set S as $\mathrm{card}(S)$. We denote a variable m uniformly distributed in the interval (a,b) as $m \sim \mathcal{U}(a,b)$.

II. Detection and isolation of sensor attacks

In this section, we consider a class of discrete-time nonlinear systems subject to sensor attacks and measurement noise. This class of systems has been considered in [25]-[28] in the attack-free case. Consider the system:

$$\begin{aligned} x^+ &= Ax + Gf(Hx) + \rho(u,y), && (1)\\ \tilde{y} &= \tilde{C}x + a + \tilde{m}, && (2)\end{aligned}$$

with state $x \in \mathbb{R}^n$, sensor measurement $\tilde{y} \in \mathbb{R}^p$, measurement noise $\tilde{m} \in \mathbb{R}^p$ satisfying $\{\tilde{m}(k)\} \in l_\infty$, matrices $A \in \mathbb{R}^{n\times n}$, $G \in \mathbb{R}^{n\times r}$, and $H \in \mathbb{R}^{r\times n}$, and attack vector $a \in \mathbb{R}^p$. If sensor $i \in \{1,\dots,p\}$ is not attacked, then the i-th component of $a(k)$ satisfies $a_i(k) = 0$ for all $k \geq 0$. Otherwise, sensor i is attacked and $a_i(k)$ is arbitrary and possibly unbounded. We denote $W \subseteq \{1,\dots,p\}$ the set of attacked sensors and thus $\mathrm{supp}(a(k)) = W$ for all $k \geq 0$. We assume the set W is fixed and unknown to us. The term $\rho(u,y)$ is a known arbitrary real-valued vector that depends on the system inputs and outputs. The state-dependent nonlinearity $f(Hx)$ is an r-dimensional vector where each entry is a function of a linear combination of the states:

$$f_i = f_i\left(\sum_{j=1}^{n} H_{ij}x_j\right), \quad i = 1,\dots,r, \tag{3}$$

with H_{ij} denotes the entries of matrix H.

Let q be the largest integer such that for each subset of sensors $J \subset \{1,\dots,p\}$ with $\mathrm{card}(J) \geq p - 2q > 0$ an observer of the form:

$$\begin{aligned} \hat{x}_J^+ =& A\hat{x}_J + Gf(H\hat{x}_J + K_J(\tilde{C}_J\hat{x}_J - \tilde{y}_J))\\ &+ L_J(\tilde{C}_J\hat{x}_J - \tilde{y}_J) + \rho(u,y), \end{aligned} \tag{4}$$

exists for $\tilde{y}_J \in \mathbb{R}^{\mathrm{card}(J)}$. Here, $\hat{x}_J \in \mathbb{R}^n$ denotes the estimate of x from $\tilde{y}_J$, and $K_J \in \mathbb{R}^{r\times \mathrm{card}(J)}$ and $L_J \in \mathbb{R}^{n\times \mathrm{card}(J)}$ are the corresponding observer matrices. The matrix $\tilde{C}_J$ is the stacking of all $\tilde{C}_i$, $i \in J$, where $\tilde{C}_i$ denotes the i-th row of $\tilde{C}$. Define the estimation error $e_J(k) := \hat{x}_J(k) - x(k)$. We assume the following.

Assumption 1 *If $a_J(k) = 0$, there exist constants $c_J > 0$, $\lambda_J \in (0,1)$, and $\gamma_J \geq 0$ satisfying:*

$$|e_J(k)| \leq c_J\lambda_J^k|e_J(0)| + \gamma_J||\tilde{m}_J||_k, \tag{5}$$

for $k \geq 0$, $e_J(0) \in \mathbb{R}^n$, and $\tilde{m}_J \in \mathbb{R}^{\mathrm{card}(J)}$, $\{\tilde{m}_J(k)\} \in l_\infty$.

Remark 1 *In this manuscript, we consider systems of the form* (1)-(2) *because under certain conditions on $f(\cdot)$, there exist tools – based on the circle-criterion – to construct observers of the form* (4) *satisfying Assumption 1. In particular, we use the result in [24], where the design method is posed as the solution of semidefinite programs.*

Assumption 2 *At most q sensors are attacked, i.e.,*

$$\mathrm{card}(W) \leq q, \tag{6}$$

and $q > 0$ is a known integer.

Assumption 3 *The bound on measurement noise is known, i.e.,*

$$||\tilde{m}||_\infty = \bar{m}, \tag{7}$$

and $\bar{m} > 0$ is a known constant.

We aim at detecting and isolating sensor attacks on system (1)-(2) for attacks and noise satisfying Assumption 2 and Assumption 3, respectively, and observers of the form (4) satisfying Assumption 1.

A. Detection of sensor attacks

We construct an observer satisfying Assumption 1 for system (1)-(2), i.e., considering all sensors, and for each subset $J \subset \{1,\dots,p\}$ of sensors with $\mathrm{card}(J) = p-q$. The obtained estimates are denoted as $\hat{x}$ and $\hat{x}_J$, respectively. Define $e = \hat{x} - x$ and let $a = 0$; then, under Assumption 1, there exist $c > 0$, $\lambda \in (0,1)$, and $\gamma \geq 0$ such that $|e(k)| \leq c\lambda^k|e(0)| + \gamma||\tilde{m}||_k$, for all $e(0) \in \mathbb{R}^n$, $k \geq 0$, and $\tilde{m} \in \mathbb{R}^p$, $\{\tilde{m}(k)\} \in l_\infty$. Because $\lambda \in (0,1)$, it can be easily verified that, for every $\epsilon > 0$, there exist k^* such that $c\lambda^k|e(0)| \leq \epsilon$, for all $k \geq k^*$, which implies $|e(k)| \leq \epsilon + \gamma||\tilde{m}||_k \leq \epsilon + \gamma\bar{m}$, for $k \geq k^*$. Also, for each subset $J \subset \{1,\dots,p\}$ with $\mathrm{card}(J) = p-q$, if $a_J = 0$, there exist $c_J > 0$, $\lambda_J \in (0,1)$, and $\gamma_J \geq 0$ such that $|e_J(k)| \leq c_J\lambda_J^k|e(0)| + \gamma_J||\tilde{m}_J||_k$, for all $e(0) \in \mathbb{R}^n$, $k \geq 0$, and $\tilde{m}_J \in \mathbb{R}^{p-q}$, $\{\tilde{m}_J(k)\} \in l_\infty$. Because $\lambda_J \in (0,1)$, there exists k_J^* such that $c_J\lambda_J^k|e(0)| \leq \epsilon$, for all $k \geq k_J^*$, and thus $|e_J(k)| \leq \epsilon + \gamma_J||m'_J||_k \leq \epsilon + \gamma_J\bar{m}$, for $k \geq k_J^*$. Let

$$k^\star := \max_{J\subset\{1,\dots,p\}:\mathrm{card}(J)=p-q} \{k^*, k_J^*\},$$

and define

$$\pi(k) := \max_{J\subset\{1,\dots,p\}:\mathrm{card}(J)=p-q} |\hat{x}(k) - \hat{x}_J(k)|. \tag{8}$$

Let $J(k) = \arg\max_{J\subset\{1,\dots,p\}:\mathrm{card}(J)=p-q} |\hat{x}(k) - \hat{x}_J(k)|$, for all $k \geq k^\star$. Then, if sensors are attack-free, i.e. $a = 0$, we have

$$\begin{aligned} \pi(k) &= |\hat{x}(k) - \hat{x}_{J(k)}(k)|\\ &= |\hat{x}(k) - x(k) + x(k) - \hat{x}_{J(k)}(k)|\\ &= |e(k) - e_{J(k)}(k)|\\ &\leq |e(k)| + |e_{J(k)}(k)|\\ &\leq 2(\epsilon + \bar{\gamma}\bar{m}), \end{aligned} \tag{9}$$

for all $k \geq k^\star$, where

$$\bar{\gamma} := \max_{J\subset\{1,\dots,p\}:\mathrm{card}(J)=p-q} \{\gamma, \gamma_J\}.$$

However, if sensors are under attack, i.e., $a \neq 0$; then, the estimates $\hat{x}(k)$ and $\hat{x}_J(k)$ in $\pi(k)$ are likely to be inconsistent and thus lead to larger $\pi(k)$ than the attack-free case. Define

$$\bar{z} := \epsilon + \bar{\gamma}\bar{m}; \tag{10}$$

then, $\bar{z}$ can be used as a threshold to detect sensor attacks for $k \geq k^\star$. However, it is still possible that for some $k \geq k^\star$ and $a_k \neq 0$, inequality (9) still holds, which would result in non detection. Then, to improve the detection rate, we perform the detection over windows of $N \in \mathbb{N}$ time-steps. That is, for each $k \in [k^\star + (i-1)N, k^\star + iN]$, $i \in \mathbb{N}$, we compute $\pi(k)$ and compare it with $\bar{z}$ for every k in the window. If there exists $k_1 \in [k^\star + (i-1)N, k^\star + iN]$, $i \in \mathbb{N}$ such that $\pi(k_1) > \bar{z}$, then we say that sensors are under attack in the i-th window. Otherwise, we say sensors are attack-free in this window. This is formally stated in Algorithm 1.

Algorithm 1 Attack Detection.

1: Design an observer satisfying Assumption 1 for system (1)-(2) and for each subset $J \subset \{1,\dots,p\}$ with $\mathrm{card}(J) = p - q$.
2: Fix the window size $N \in \mathbb{N}$.
3: Calculate $\bar{z}$ as in (10).
4: For $i \in \mathbb{Z}_{>0}$, calculate $\pi(k)$ for $k \in [k^\star + (i-1)N, k^\star + iN - 1]$.
5: For $i \in \mathbb{Z}_{>0}$, if $\exists\ k_1 \in [k^\star + (i-1)N, k^\star + iN - 1]$ such that $\pi(k_1) > \bar{z}$, then sensor attacks occurs in the i-th window, and

$$detection(i) = 1;$$

otherwise, sensors are attack-free in the i-th window, and

$$detection(i) = 0.$$

6: Return $detection(i)$

Because our knowledge of $||m||_\infty$ might be conservative, we consider the case when the actual bound on measurement noise is smaller than $\bar{m}$, i.e., $||m||_\infty = \tau\bar{m}$ and $\tau \in (0,1)$. We give a sufficient condition under which sensor attacks cannot be detected by Algorithm 1 in the i-th time window for a given $N > 0$.

Proposition 1 *Given a time window length $N > 0$, if*

$$||a||_{k^\star+iN-1} \leq (1-\tau)\bar{m}; \tag{11}$$

where $\tau \in (0,1)$, then, $\pi(k) \leq \bar{z}$ for all $k \in [k^\star + (i-1)N, k^\star + iN - 1]$ and sensor attacks cannot be detected by Algorithm 1 in the i-th time window.

Proof: For a given time window length $N > 0$, sensor attacks cannot be detected by Algorithm 1 in the i-th time window for $a \neq 0$, if we have

$$\pi(k) \leq \bar{z}, \tag{12}$$

for all $k \in [k^\star + (i-1)N, k^\star + iN - 1]$. For $a \neq 0$, we have

$$\begin{aligned}\pi(k) \leq& |e(k) + |e_J(k)| \\ \leq& 2\epsilon + \gamma(||\tilde{m}||_{k^\star+iN-1} + ||a||_{k^\star+iN-1}) \\ &+ \gamma_J(||\tilde{m}||_{k^\star+iN-1} + ||a_J||_{k^\star+iN-1}) \\ \leq& 2(\epsilon + \bar{\gamma}(||\tilde{m}||_\infty + ||a||_{k^\star+iN-1}) \\ \leq& 2(\epsilon + \bar{\gamma}(\tau\bar{m} + ||a||_{k^\star+iN-1})),\end{aligned}$$

for all $k \in [k^\star + (i-1)N, k^\star + iN - 1]$. It follows that the inequality (12) is satisfied for a satisfying (11) for all $k \in [k^\star + (i-1)N, k^\star + iN - 1]$ and thus sensor attacks cannot be detected by Algorithm 1 in the i-th time window. ■

Next, we give a sufficient condition under which sensor attacks can always be detected by Algorithm 1 in the i-th time window for a given $N > 0$.

Proposition 2 *For a given time window length $N > 0$, if there exist $k_1 \in [k^\star + (i-1)N, k^\star + iN - 1]$ such that*

$$|e(k_1)| > 3(\epsilon + \bar{\gamma}\bar{m}); \tag{13}$$

then, $\pi(k_1) > \bar{z}$ and thus sensor attacks can be detected by Algorithm 1 in the i-th time window.

Proof: For a given time window length $N > 0$, sensor attacks can be detected by Algorithm 1 in the i-th time window for $a \neq 0$, if there exist $k_1 \in [k^\star + (i-1)N, k^\star + iN - 1]$ such that $\pi(k_1) > \bar{z}$. Since there are at most q sensors under attack, we know there exist at least one $\bar{I} \subset \{1,\dots,p\}$ with $\mathrm{card}(\bar{I}) = p - q$ such that $a_{\bar{I}} = 0$, and

$$|e_{\bar{I}}(k)| \leq \epsilon + \gamma_{\bar{I}}||\tilde{m}_{\bar{I}}||_k, \tag{14}$$

for $k \geq k^\star$. From (8), we know $\pi(k) \geq |e(k) - e_{\bar{I}}(k)|$ for $k \geq k^\star$. If (13) holds, then

$$\begin{aligned}\pi(k_1) &\geq ||e(k_1)| - |e_{\bar{I}}(k_1)|| \\ &> 3(\epsilon + \bar{\gamma}\bar{m}) - \epsilon - \gamma_{\bar{I}}||\tilde{m}_{\bar{I}}||_{k_1} \\ &> 2(\epsilon + \bar{\gamma}\bar{m}),\end{aligned} \tag{15}$$

which implies sensor attacks can be detected by Algorithm 1 in the i-th time window. ■

B. Isolation of sensor attacks

To perform the isolation, we construct an observer satisfying Assumption 1 for each subset $J \subset \{1,\dots,p\}$ of sensors with $\mathrm{card}(J) = p - q$ and each subset $S \subset \{1,\dots,p\}$ of sensors with $\mathrm{card}(S) = p - 2q$. Hence, by Assumption 1, for $a_S(k) = 0$, there exist $c_S > 0$, $\lambda_S \in (0,1)$, and $\gamma_S \geq 0$ satisfying

$$|e_S(k)| \leq c_S\lambda_S^k|e(0)| + \gamma_S||\tilde{m}_S||_k, \tag{16}$$

for all $e(0) \in \mathbb{R}^n$ and $k \geq 0$. Note that, because $\lambda_S \in (0,1)$, there always exist k_S^* such that $c_S\lambda_S^k|e(0)| \leq \epsilon$, for any $\epsilon > 0$ and $k \geq k_S^*$. Define $\bar{k}^* := \max_{J,S}\{k_J^*, k_S^*\}$. For each subset J with $\mathrm{card}(J) = p - q$, define $\pi_J(k)$ as

$$\pi_J(k) := \max_{S\subset J:\mathrm{card}(S)=p-2q} |\hat{x}_J(k) - \hat{x}_S(k)|. \tag{17}$$

Since there are at most q sensors under attack, we know there exist at least one $\bar{I} \subset \{1,\dots,p\}$ with $\mathrm{card}(\bar{I}) = p-q$ such that $a_{\bar{I}} = 0$ and (14) is satisfied. Define

$$\begin{aligned}\pi_{\bar{I}}(k) &:= \max_{S\subset\bar{I}} |\hat{x}_{\bar{I}}(k) - \hat{x}_S(k)| \\ &= \max_{S\subset\bar{I}} |\hat{x}_{\bar{I}}(k) - x(k) + x(k) - \hat{x}_S(k)| \\ &\le |e_{\bar{I}}(k)| + \max_{S\subset\bar{I}} |e_S(k)|. \end{aligned} \tag{18}$$

From (14) and (16), we obtain $\pi_{\bar{I}}(k) \le 2(\epsilon + \gamma'_{\bar{I}} ||\tilde{m}_{\bar{I}}||_k)$, for all $k \ge \bar{k}^*$, where

$$\gamma'_{\bar{I}} := \max_{S\subset\bar{I}:\mathrm{card}(S)=p-2q} \{\gamma_{\bar{I}}, \gamma_S\}.$$

However, if the subset J of sensors is under attack, i.e., $a_J \neq 0$, then $\hat{x}_J(k)$ and $\hat{x}_S(k)$ in $\pi_J(k)$ are more inconsistent and might produce larger $\pi_J(k)$. Define

$$\bar{z}_J = 2(\epsilon + \gamma'_J \bar{m}), \tag{19}$$

for each $J \subset \{1,\dots,p\}$ with $\mathrm{card}(J) = p-q$, where

$$\gamma'_J := \max_{S\subset J:\mathrm{card}(S)=p-2q} \{\gamma_J, \gamma_S\};$$

then, $\bar{z}_J$ can be used as a threshold to isolate attacked sensors. For all $k \ge \bar{k}^*$, we select out all the subsets $J \subset \{1,\dots,p\}$ with $\mathrm{card}(J) = p-q$ that satisfy

$$\pi_J(k) \le \bar{z}_J. \tag{20}$$

Denote as $\bar{W}(k)$ the set of sensors that we regard as attack-free at time k. Then, $\bar{W}(k)$ is given as the union of all subsets J such that (20) holds:

$$\bar{W}(k) := \bigcup_{J\subset\{1,\dots,p\}:\mathrm{card}(J)=p-q, \pi_J(k)\le\bar{z}_J} J. \tag{21}$$

Thus, the set $\{1,\dots,p\} \setminus \bar{W}(k)$ is isolated as the set of attacked sensors at time k. However, note that it is still possible that for some $k \ge \bar{k}^*$ and some $J \subset \{1,\dots,p\}$ with $\mathrm{card}(J) = p-q$, $a_J(k) \neq 0$ but (20) still holds. This implies that $J \subset \bar{W}(k)$ even if $a_J \neq 0$ and would result in wrong isolation. Therefore, we perform the isolation over windows of $N \in \mathbb{N}$ time-steps. That is, for each $k \in [\bar{k}^* + (i-1)N, \bar{k}^* + iN]$, $i \in \mathbb{N}$, we compute and collect $\bar{W}(k)$ for every k in the window and select the subset J with $\mathrm{card}(J) \ge p-q$ that is equal to $\bar{W}(k)$ most often in the i-th window. We denote this J as $J(i)$. Then, we select $\{1,\dots,p\} \setminus J(i)$ as the set of sensors under attack in the i-th window. This is formally stated in Algorithm 2.

Next, we give a sufficient condition under which none of the attacked sensors can be isolated by Algorithm 2 in the i-th time window for a given $N > 0$ when $||m||_\infty = \tau \cdot \bar{m}$ where $\tau \in (0,1)$.

Proposition 3 *Given a time window length $N > 0$, if*

$$||a||_{\bar{k}^*+iN-1} \le (1-\tau)\bar{m}; \tag{22}$$

where $\tau \in (0,1)$, then, for all $J \subset \{1,\dots,p\}$ with $\mathrm{card}(J) = p-q$ and $a_J \neq 0$, $\pi_J(k) \le \bar{z}_J$ for all $k \in \left[\bar{k}^* + (i-1)N, \bar{k}^* + iN - 1\right]$ and none of attacked sensors can be isolated by Algorithm 2 in the i-th time window.

Algorithm 2 Attack Isolation.

1: Design an observer satisfying Assumption 1 for each subset $J \subset \{1,\dots,p\}$ with $\mathrm{card}(J) = p-q$ and each subset $S \subset \{1,\dots,p\}$ with $\mathrm{card}(S) = p-2q$.
2: Intialize the counter variable $n_J(i) = 0$ for all J with $\mathrm{card}(J) \ge p-q$ and all $i \in \mathbb{Z}_{>0}$.
3: Calculate $\bar{z}_J$ for each J with $\mathrm{card}(J) = p-q$ as (19).
4: For $i \in \mathbb{Z}_{>0}$ and $\forall k \in \left[\bar{k}^* + (i-1)N, \bar{k}^* + iN - 1\right]$, calculate $\pi_J(k)$, $\forall J$ with $\mathrm{card}(J) = p-q$ as follows:
$$\pi_J(k) = \max_{S\subset J:\mathrm{card}(S)=p-2q} |\hat{x}_J(k) - \hat{x}_S(k)|.$$
5: For all $k \in \left[\bar{k}^* + (i-1)N, \bar{k}^* + iN - 1\right]$, take the union of all the subsets J such that $\pi_J(k) \le \bar{z}_J$:
$$\bar{W}(k) = \bigcup_{J\subset\{1,\dots,p\}:\mathrm{card}(J)=p-q, \pi_J(k)\le\bar{z}_J} J,$$
6: For all $k \in \left[\bar{k}^* + (i-1)N, \bar{k}^* + iN - 1\right]$, if $\bar{W}(k) = J$ for some J with $\mathrm{card}(J) \ge p-q$, then update its corresponding counter variable as follows:
$$n_J(i) = n_J(i) + 1.$$
7: For all $i \in \mathbb{Z}_{>0}$, select the subset J with $\mathrm{card}(J) \ge p-q$ that is equal to $\bar{W}(k)$ most often, i.e.,
$$J(i) = \arg\max_{J\in\{1,\dots,p\}:\mathrm{card}(J)\ge p-q} n_J(i).$$
8: For all $i \in \mathbb{Z}_{>0}$, the set of sensors potentially under attack is given as:
$$\tilde{A}(i) = \{1,\dots,p\} \setminus J(i).$$
9: For all $i \in \mathbb{Z}_{>0}$, return $\tilde{A}(i)$.

Proof: For given time window length $N > 0$, none of attacked sensors can be isolated by Algorithm 2 in the i-th window if $\forall J \subset \{1,\dots,p\}$ with $\mathrm{card}(J) = p-q$ and $a_J \neq 0$, we have $\pi_J(k) \le \bar{z}_J$ for all $k \in \left[\bar{k}^* + (i-1)N, \bar{k}^* + iN - 1\right]$. For all $k \in \left[\bar{k}^* + (i-1)N, \bar{k}^* + iN - 1\right]$ and $a_J \neq 0$, we have

$$\begin{aligned}\pi_J(k) \le& |e_J(k) + |e_S(k)| \\ \le& 2\epsilon + \gamma_J(||\tilde{m}_J||_\infty + ||a_J||_\infty) \\ &+ \gamma_S(||\tilde{m}_S||_{\bar{k}^*+iN-1} + ||a_S||_{\bar{k}^*+iN-1}) \\ \le& 2\epsilon + 2\gamma'_J(||\tilde{m}_J||_{\bar{k}^*+iN-1} + ||a_J||_{\bar{k}^*+iN-1}) \\ \le& 2(\epsilon + \gamma'_J(\tau\bar{m} + ||a||_{\bar{k}^*+iN-1})).\end{aligned}$$

If (22) holds, then $\pi_J(k) \le \bar{z}_J$ for all $J \subset \{1,\dots,p\}$ with $\mathrm{card}(J) = p-q$ and all $k \in \left[\bar{k}^* + (i-1)N, \bar{k}^* + iN - 1\right]$. Then, none of attacked sensors can be isolated by Algorithm 2 in the i-th time window. ■

Next, we give a sufficient condition under which all of attacked sensors can be isolated by Algorithm 2 in the i-th time window for a given time window length $N > 0$.

Proposition 4 *Given a time window length $N > 0$, if for all $J \subset \{1, \dots, p\}$ with $\text{card}(J) = p - q$ and $a_J \neq 0$, we have*

$$|e_J(k)| > 3(\epsilon + \gamma'_J \bar{m}), \tag{23}$$

for at least $N/2$ time-steps in the i-th time window, then for all $J \subset \{1, \dots, p\}$ with $\text{card}(J) = p - q$ and $a_J \neq 0$, we have $\pi_J(k) > \bar{z}_J$ for at least $N/2$ time-steps in the i-th time window, and all of attacked sensors can be isolated by Algorithm 2 in the i-th time window.

Proof: Since there are at most q sensors under attack, for each subset J with $\text{card}(J) = p - q$ we know there exist at least one $\bar{S} \subset J$ with $\text{card}(\bar{S}) = p - 2q$ such that $a_{\bar{S}} = 0$, and $|e_{\bar{S}}| \leq \epsilon + \gamma_{\bar{S}} ||\tilde{m}_{\bar{S}}||_k$, for all $k \in \left[\bar{k}^* + (i-1)N, \bar{k}^* + iN - 1\right]$. By construction of (17), it is satisfied that

$$\begin{aligned} \pi_J(k) &= \max_{S \subset J \, \text{card}(S) = p - 2q} |\hat{x}_J(k) - \hat{x}_S(k)| \\ &\geq |e_J(k) - e_{\bar{S}}(k)|. \end{aligned}$$

for all $k \geq \bar{k}^*$. If (23) holds at least $N/2$ time-steps in the i-th time window, then from triangle inequality, for all $J \subset \{1, \dots, p\}$ with $\text{card}(J) = p - q$ and $a_J \neq 0$, we have

$$\begin{aligned} \pi_J(k) &\geq ||e_J(k)| - |e_{\bar{S}}(k)|| \\ &> 3(\epsilon + \gamma'_J \bar{m}) - \epsilon - \gamma_{\bar{S}} ||\tilde{m}_{\bar{S}}||_k \\ &> 2(\epsilon + \gamma'_J \bar{m}), \end{aligned} \tag{24}$$

for at least $N/2$ time-steps in the i-th time window, which implies all of attacked sensors can be isolated by Algorithm 2 in the i-th time window. ■

Remark 2 *The performance of Algorithm 2 can be arbitrarily improved by increasing the length of the time window N at the price of increasing the time needed for isolation.*

Example 1 Consider the discrete-time nonlinear system subject to measurement noise and sensor attacks:

$$x^+ = \begin{bmatrix} 1 & \delta \\ 0 & 1 \end{bmatrix} x + \begin{bmatrix} \frac{1}{2}\delta\alpha \sin(x_1 + x_2) \\ \delta\alpha \sin(x_1 + x_2) \end{bmatrix} + \begin{bmatrix} \delta u \\ \delta u \end{bmatrix}, \tag{25}$$

$$\tilde{y} = \begin{bmatrix} 3 & 0.3 \\ 3 & 0.6 \\ 6 & 0.9 \\ 1.2 & 12 \end{bmatrix} x + a + \tilde{m}. \tag{26}$$

with $\delta = 0.1$, $\alpha = 1$, and $\tilde{m}_i \sim \mathcal{U}(-0.5, 0.5)$ for $i \in \{1, 2, 3, 4\}$. Using the design method proposed in [24], we find that circle-criterion observers of the form (4) satisfying Assumption 1 exist for each subset $J \subset \{1, 2, 3, 4\}$ with $\text{card}(J) \geq 1$. Since $p = 4$, by Assumption 2, the maximum number of attacks is $q = 1$. We design a circle-criterion observer for the whole system and for each $J \subset \{1, 2, 3, 4\}$ with $\text{card}(J) = 3$ and each $S \subset \{1, 2, 3, 4\}$ with $\text{card}(S) = 2$. Therefore, in total, $\binom{4}{3} + \binom{4}{2} + 1 = 11$ observers are designed. We obtain their ISS gains by montecarlo simulations. Theses eleven observers are initialized at $\hat{x}(0) = x(0)$ and $x_1(0), x_2(0)$ are randomly selected from a standard normal distribution; thus, $\epsilon = 0$. We let $N = 50, 100, 200$ and evaluate Algorithm 1 and Algorithm 2 for 1000 time-steps. For Algorithm 1, we let $W = \{2\}$, which means the 2-nd sensor is under attack, and $a_2 \sim \mathcal{U}(-c, c)$ with c given by 0.7 and 1. We run Algorithm 1 with $\binom{4}{3} + 1 = 5$ observers. The detection results are shown in Figures 1-2. For Algorithm 2, we let $W = \{3\}$ and $a_3 \sim \mathcal{U}(-d, d)$ with d given by 2 and 5. We run Algorithm 2 with $\binom{4}{3} + \binom{4}{2} = 10$ observers. We say sensor 0 is under attack in the i-th window when $\tilde{A}_i = \emptyset$. The isolation results are shown below in Figures 3-4.

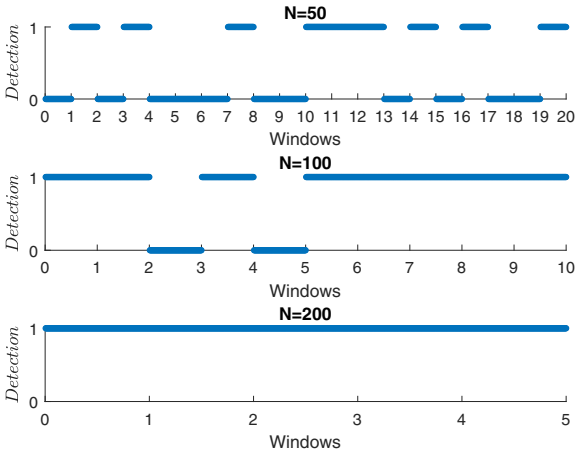


Fig. 1. Attack detection, $a_2 \sim \mathcal{U}(-0.7, 0.7)$.

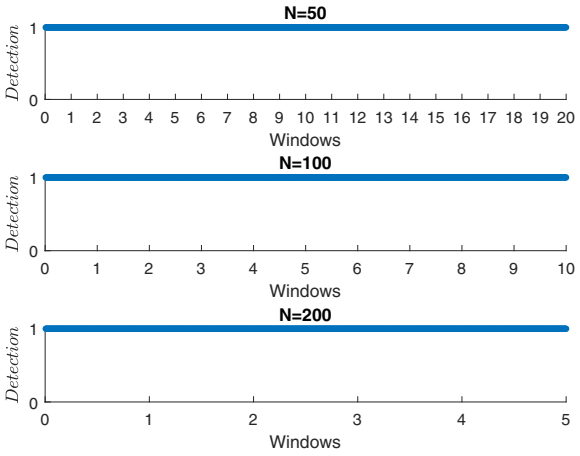


Fig. 2. Attack detection, $a_2 \sim \mathcal{U}(-1, 1)$.

III. Conclusion

Assuming that a sufficiently small subset of sensors is subject to additive false data injection attacks, we have proposed two algorithms for detecting and isolating sensor attacks for a class of discrete-time nonlinear systems subject to measurement noise using a multi-observer approach. We

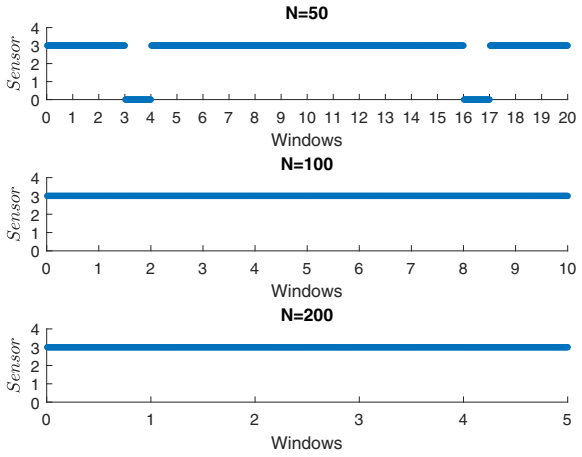


Fig. 3. Attack isolation, $a_3 \sim \mathcal{U}(-2, 2)$.

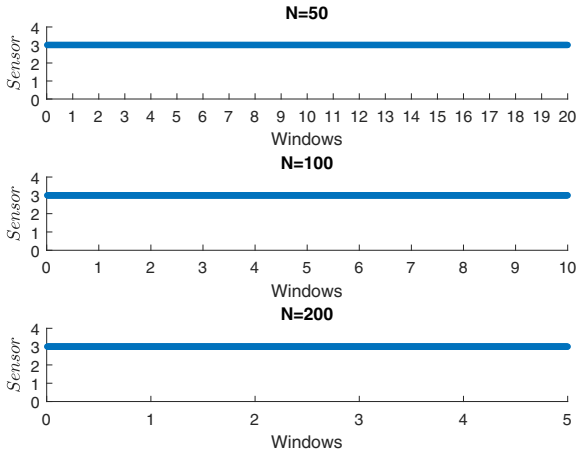


Fig. 4. Attack isolation, $a_3 \sim \mathcal{U}(-5, 5)$.

have provided simulations results to illustrate the performance of the proposed algorithms. The performance of our algorithms can be improved by increasing the length of the time window N at the price of increasing the time needed for detection and isolation.

REFERENCES

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

[2] M. Massoumnia, G. C. Verghese, and A. S. Willsky, "Failure detection and identification in linear time-invariant systems," *Technology*, no. July, 1986.

[3] Y. Mo and B. Sinopoli, "Resilient detection in the presence of integrity attacks," *IEEE Transactions on Signal Processing*, vol. 62, no. 1, pp. 31–43, 2014.

[4] K. G. Vamvoudakis, J. P. Hespanha, B. Sinopoli, and Y. Mo, "Detection in adversarial environments," *IEEE Transactions on Automatic Control*, vol. 59, no. 12, pp. 3209–3223, 2015.

[5] M. S. Chong and M. Kuijper, "Characterising the vulnerability of linear control systems under sensor attacks using a system's security index," in *IEEE 55th Conference on Decision and Control (CDC)*, pp. 5906–5911, 2016.

[6] K. G. Vamvoudakis, J. P. Hespanha, B. Sinopoli, and Y. Mo, "Adversarial detection as a zero-sum game," in *IEEE 51st Conference on Decision and Control (CDC)*, pp. 7133–7138, 2012.

[7] Y. Shoukry, P. Nuzzo, A. Puggelli, A. L. Sangiovanni-Vincentelli, S. Seshia, and P. Tabuada, "Secure state estimation for cyber physical systems under sensor attacks: a satisfiability modulo theory approach," *IEEE Transactions on Automatic Control*, vol. 62, pp. 4917 – 4932.

[8] S. Z. Yong, M. Zhu, and E. Frazzoli, "Resilient state estimation against switching attacks on stochastic cyber-physical systems," in *IEEE 54th Conference on Decision and Control (CDC)*, pp. 5162–5169, 2015.

[9] J. Park, J. Weimer, and I. Lee, "Sensor attack detection in the presence of transient faults," *6th International Conference on Cyber-Physical Systems*, no. April, pp. 1–10, 2015.

[10] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Transactions on Information and System Security*, vol. 14, no. 1, pp. 21–32, 2009.

[11] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "Revealing stealthy attacks in control systems," *2012 50th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2012*, pp. 1806–1813, 2012.

[12] C. Murguia and J. Ruths, "Characterization of a CUSUM model-based sensor attack detector," in *2016 IEEE 55th Conference on Decision and Control, CDC 2016*, pp. 1303–1309, 2016.

[13] V. S. Dolk, P. Tesi, C. D. Persis, and W. P. M. H. Heemels, "Event-triggered control systems under denial-of-service attacks," *IEEE Transactions on Control of Network Systems*, vol. 4, pp. 93–105, 2017.

[14] N. Hashemil, C. Murguia, and J. Ruths, "A comparison of stealthy sensor attacks on control systems," in *proceedings of the American Control Conference (ACC)*, 2017.

[15] F. Pasqualetti, F. Dorfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 58, pp. 2715–2729, 2013.

[16] C. Murguia and J. Ruths, "On reachable sets of hidden cps sensor attacks," in *proceedings of the American Control Conference (ACC)*, 2017.

[17] J. Giraldo, A. Cardenas, and N. Quijano, "Integrity attacks on real-time pricing in smart grids: Impact and countermeasures," *IEEE Transactions on Smart Grid*, 2016.

[18] C. Murguia and J. Ruths, "Cusum and chi-squared attack detection of compromised sensors," in *proceedings of the IEEE Multi-Conference on Systems and Control (MSC)*, 2016.

[19] S. H. Kafash, J. Giraldo, C. Murguia, A. A. Cardenas, and J. Ruths, "Constraining attacker capabilities through actuator saturation," in *proceedings of the American Control Conference (ACC)*, 2017.

[20] C. Murguia, N. van de Wouw, and J. Ruths, "Reachable sets of hidden cps sensor attacks: Analysis and synthesis tools," in *proceedings of the IFAC World Congress*, 2016.

[21] C. Murguia, I. Shames, F. Farokhi, and D. Nešić, "On privacy of quantized sensor measurements through additive noise," in *proceedings of the 57th IEEE Conference on Decision and Control (CDC)*, 2018.

[22] C. Murguia, I. Shames, J. Ruths, and D. Nešić, "Security metrics of networked control systems under sensor attacks (extended preprint)," in *arXiv:1809.01808*, 2018.

[23] J. Kim, C. Lee, H. Shim, Y. Eun, and J. H. Seo, "Detection of sensor attack and resilient state estimation for uniformly observable nonlinear systems," *IEEE 55th Conference on Decision and Control, CDC*, 2016.

[24] Y. Tianci, M. Carlos, M. Kuijper, and D. Nešić, "A robust circle-criterion observer-based estimator for discrete-time nonlinear systems in the presence of sensor attacks," *IEEE 57th Conference on Decision and Control (CDC), accepted*, 2018.

[25] S. Ibrir, "Circle-criterion approach to discrete-time nonlinear observer design," *Automatica*, vol. 43, no. 8, pp. 1432–1441, 2007.

[26] S. Sundaram, "State and unknown input observers for discrete-time nonlinear systems," in *IEEE 55th Conference on Decision and Control (CDC)*, pp. 7111–7116, 2016.

[27] A. Zemouche and M. Boutayeb, "Observer Design for Lipschitz Nonlinear Systems: The Discrete-Time Case," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 53, no. 8, pp. 777–781, 2006.

[28] M. Abbaszadeh and H. J. Marquez, "Robust H_∞ observer design for sampled-data Lipschitz nonlinear systems with exact and Euler approximate models," *Automatica*, vol. 44, no. 3, pp. 799–806, 2008.