Linear Cyber-Physical System Security – Detection and Correction of Adversarial Attacks

Zhanghan Tang

Submitted in total fulfilment of the requirements of the degree of

Doctor of Philosophy

Department of Electrical and Electronic Engineering
THE UNIVERSITY OF MELBOURNE

November 2019

Abstract

Malicious attacks on Cyber-Physical Systems (CPS) may cause significant damage to the targeted system. In this thesis, we address the problem of attack detection as well as attack correction for multi-input multi-output discrete-time linear time-invariant dynamical systems when the attacker can inject signals (additively) to sensors and actuator signals. We consider the cases of sensor only attacks, actuator only attacks and consider the case when both sensors and actuators are attacked. We also consider the case when we have prior knowledge about a specific subset of sensors and/or actuators that is not accessible by the attacker. For each attack scenario (sensor and/or actuator attacks with or without prior knowledge), we present attack detection and correction methods.

In this thesis, we present novel methods for solving the problems of attack detection and correction using the notion of 'security index' for various attack scenarios. The security index is a system parameter which characterises the system's vulnerability against different types of attacks. In addition, it provides a quantitative measurement for measurement redundancy. In this thesis, we first present the security index as a representation-free system parameter. To achieve this, we use a behavioural approach as our starting point.

To solve the detection and correction problems, we use a variety of system representations, but mostly Input/Latent/Output image representations.

Declaration

I, Zhanghan TANG, declare that this thesis titled 'Linear Cyber-Physical System Security – Detection and Correction of Adversarial Attacks' and the work presented in it are my own. I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University.
2. Where any part of this thesis has previously been submitted for a degree or any other qualifications, this has been clearly stated.
3. Where I have consulted the published work of others, this is always clearly attributed.
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
5. I have acknowledged all main sources of help.
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.
7. The thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Zhanghan Tang, November 2019

Acknowledgements

My PhD candidature would not have been successful without the strong support for many amazing people. I would like to give my sincere thanks to all those people who have helped me during this remarkable journey in my life.

First of all, I would like to extend my gratitude to my principle supervisor Prof. Margreta Kuijper and my co-supervisor Prof. Iven Mareels, for their excellent supervision and support during my PhD candidature. Their great enthusiasm for the research has been a source of great inspiration to me. Their careful reviews and sharp criticisms are essential in the improvement of this thesis. I am grateful to them for constantly expanding my knowledge and academic horizons.

Secondly, many thanks go to the academic and staff of the Department of Electrical and Electronic Engineering in the University of Melbourne. Our deputy head Prof. Girish Nair, Department Administrator Lyn Bunchanan, Natasha Baxter, were kind enough to provide powerful resources and environmental support throughout my PhD candidature. I would also like to thank the chair of my advisory committee Prof. Marimuthu Palaniswami and my committee member Prof. Christopher Leckie for their valuable comments and suggestions.

Moreover, I would like to give my sincere thanks to all the friends and colleagues, including Dr. Michelle Chong from KTH Royal Institute of Technology, Tianci Yang, Dr. Zhiyang Ju, Dr. Zibo Miao, Dr. Zhe Wang, Dr. Shawn Cheng, Jeanne Lefevre, Ali, Asif, Majid. Thank you all for sharing my journey and being supportive in more ways than one.

I would also like to thank the Authentication for the Future Internet of Things Forum (AuthIoT 2018) hosted by Deakin University in November 2018, where my supervisor

and I met Dr. Joanne Hall from RMIT university and take the initiative for a brief discussion regarding CPS security and its engineering application which then forms the Chapter 6 of this thesis.

Last but not least, I would like to express my deepest gratitude to my beloved family for their endless love, dedication and support. I will always member it, treasure it, from the bottom of my heart. The emotional support from my significant other half Naixin Geng gives me great courage and confidence to achieve the completion of my PhD candidature and this thesis. A special thanks to my parents Jun Tang and Kerong Zhang; I appreciate everything that you have done to shape my life and support my education.

Preface

The outcomes of the thesis are published in the following journal and conferences. For all the publications, the first author Z. Tang contributed greater than 50% of the content of each publication and is the primary author. The first author Z. Tang was primarily responsible for the planning, execution and preparation of the work for publication. However, the first author Z. Tang benefited from his supervisors and colleagues from group meeting sessions or email communications where they provided technical comments and guidance. The publications and the contributions of each author are listed as follows:

- Z.Tang, M. Kuijper, M. Chong, I. Mareels, C. Leckie. Linear system security – detection and correction of adversarial sensor attacks in the noise-free case. Automatica, 101 (2019), 53-59.

  First author: problem formulation; finding a suitable approach based on literature review; completing mathematical theorems and proofs; construct simulations; paper writing; revision.

  Second author: supervision; providing technical comments; proofreading; paper polishing.

  Third and Fourth author: providing technical comments; proof reading.

  Fifth author: consultation.

- Z.Tang, M. Kuijper, M. Chong, I. Mareels, C. Leckie. Attack correction for noise-free linear systems subject to sensor attacks. In 23rd International Symposium on Mathematical Theory of Networks and Systems (MTNS 2017), 17-21.

  First author: problem formulation; finding a suitable approach based on literature

review; completing mathematical theorems and proofs; constructing simulations; paper writing; revision; presenting the work at the conference.

Second author: supervision; providing technical comments; proofreading; paper polishing.

Third and Fourth author: providing technical comments; proof reading.

Fifth author: consultation.

- Z.Tang, M. Kuijper, M. Chong, I. Mareels, C. Leckie. Sensor attack correction for linear systems with known input. IFAC workshop on Distributed Estimation and Control in Networked Systems (7th NECSYS 2018), available at IFAC-PaperOnLine, vol. 51, no.23, 206-211.

  First author: problem formulation; finding a suitable approach based on literature review; completing mathematical theorems and proofs; constructing simulations; paper writing; revision.

  Second author: supervision; providing technical comments; proofreading; paper polishing.

  Third author: providing technical comments; presenting the work at the conference.

  Fourth author: providing technical comments; proof reading.

  Fifth author: consultation.

Financial support provided by the Australian Government under Research Training Scheme and Melbourne School of Engineering Travel Grant are gratefully acknowledged.

Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Cyber-Physical System Security

The terminology 'Cyber-Physical-System' (CPS) was coined by Helen Gill at the National Science Foundation in the United States in 2006. Notably, the term CPS refers to a new generation of systems with integrated computational and physical capabilities that is capable of interacting with humans through many new modalities [6]. Examples of the CPSs include vehicular systems and transportation, smart homes and buildings, power systems, networking systems, robotics, and so on. Not long after the concept of CPS being proposed, research topics in CPS also began to emerge. One of the research topics is CPS security. The roots of CPS security date back to the concepts of fault detection and isolation (FDI) in the control community. The objective of fault detection is to determine the existence of faults, while the objective of fault isolation is to pinpoint the type of fault as well as its location. The work [34] presents a survey of the various model-based FDI methods developed in the last decade. Publications on FDI for linear systems dates back to 1966 when S. Seshu and R. Waxman [55] proposed a procedure for generating a set of test points to isolate faults for linear systems. In 1980, E. Y. Chow [17] addressed the problem of FDI using a residual generation process. Chow and Willsky's work [16] builds the foundation of FDI for linear systems.

From 2006, with the advent of the concept of CPS and the appearance of several severe CPS security incidents caused by adversarial attackers, the topic of attack detection and attack correction gained prominence. Although attacks on a system may exhibit similar traits as faults, significant differences do exist between faults and attacks. Faults

usually entail inherent patterns based on the type of fault being considered. For example, a loss of conductivity in a wire usually implies the current drops in the circuit. Unlike faults, attacks are mounted by an attacker with malicious intent, meaning the attack signal injected by the attacker commits specific acts that will result in biasing the system to achieve specific malicious behaviour.

The essential difference between a fault and an attack can be summarised as follows. A fault may create damage but there was no intent to cause damage, whereas an attack has an intent its motivation was to create damage or to mislead. Another important feature in a malicious, intentful attack is that, the attacker may know the system model and thus, can mimic the behaviour of the system at a certain level.

The work of A. W. Werth [70] explains the similarities and differences between a fault and an attack for CPS systems. The general definition of a fault is: 'unintended deviations from the normal behaviour of the plant or its instrument'. These faults could lead to damage to equipment, errors in the network or certain malfunctions in software. In this sense, attacks will exhibit similar characteristics to faults. However, [70] further explains that an attack is an action which undermines the security of a system for malicious purposes, that is, the attacker has shown intent in launching the attack, or has a definitive purpose. Attacks may include actions such as, jamming communication channels by sending superfluous data, or intentionally sending incorrect information. On the other hand, a fault has no purpose or intent. It is a random act, due to misbehaviour of the plant, perhaps due to manufacturing errors, or ageing, or being operated outside of its intended operational limits.

1.2 Motivation

Real-world attacks on CPS have occurred in the past decade and caused significant damage to the targeted systems in some cases, both economically and often, threatening people's safety.

Based on the information from the Congressional Research Service Report [19] in 2011, which investigated the economic impact of malicious attacks to a CPS, the target

firm may lose up to 5% in stock market valuation the day after an attack has been revealed to the market. In 2003, a negative economic impact in the USA was estimated around 250 billion.

A CPS attack can also threaten people's safety. For example, in January 2008, a 14 year-old teenager in Poland used a modified remote controller to attack the track switch system of a city tram, which caused twelve injuries. In 2000, a former employee managed to hack into the control system of the Maroochy water services system in Queensland, Australia, causing thousand tons of sewage water drowned the surrounding parks, houses and a hotel, see [59] for more detail. It is also conceivable that a modern autonomous vehicle may be hijacked using sensor spoofing [67].

How to reduce the impact of such malicious attacks is the principle motivation of this work. In general, there are two ways of curtailing the impact of an attack: robust system design and signal reconstruction. Robust system design [7, 43] aims to reduce the impact of an attack from the design perspective of a system, while signal reconstruction seeks methods to reconstruct the attack-free signals on the premise of the attack signal being detected. We follow the latter approach. More specifically, given a system that is potentially attacked, we first propose an attack detection method in order to identify whether an attack has occurred. We then propose an attack correction method to reconstruct the attack-free input/output signals thus limiting the impact of such attacks. In this thesis, we speak about detection and not prediction and that in reconstructing attack free signals, the methods do take time. Put succinctly, delay is inevitable in detection/correction. The attack still has influence, but only over the time it took to detect and neutralise the attack signal.

1.3 Thesis Outline

The thesis is organised as follows.

In Chapter 2, we review the relevant literature in the area of CPS security, along with some classic fault detection and identification literature. We make it clear that we are focusing on linear time-invariant discrete time systems throughout the thesis. We also

review relevant concepts and representations that will be used in later chapters.

In Chapter 3, the problem of sensor only attack detection and attack correction for linear dynamical systems is discussed. In this section, we first recall the concepts around a system's kernel representation. This is the starting point for our approach. In particular, we consider two different scenarios: we first consider the general case when the attacker potentially has access to all sensor signals. We then consider the case where we know which subset of sensors is not attacked. We build on the previous works of [15] and [14] and recall the concept of (sensor) security index, which characterises the vulnerability of a dynamical system against sensor attack. Thereafter, attack detection and correction algorithms are proposed and developed. The proposed algorithms are guaranteed to achieve sensor attack detection and correction under certain assumptions with respect to the attack signal.

In Chapter 4, dynamical systems under actuator only attack are considered. The approach followed in this chapter is similar to the one we followed in Chapter 3. Analogous to the sensor attack case, we propose a new concept of actuator security index. Attack detection and correction algorithms are developed, first for general systems without prior knowledge as to which of the actuators could be attacked. Next, we consider the case where we know that some actuators are attack-free.

Based on our previous knowledge regarding sensor only attack and actuator only attack, in Chapter 5, we consider the situation when a system may experience both actuator and sensor attacks simultaneously. For the purpose of attack correction, we aim to achieve attack correction when (potentially) all actuators are attacked together with some (but not all) sensors being attacked. In order to achieve our aim, the strong observability plays a vital role in this chapter. For this reason, in Chapter 5, we assume that the systems (or specific subsystems) are strongly observable.

Chapter 3, 4 and 5 assume that all signals are measured with total accuracy and that the measurement process is noise free; these ideal assumptions provide conceptual approaches regarding CPS security. However, in reality, sensors have finite resolution. Actuators have finite resolution as well as a finite dynamic range. Chapter 6 commences the journey to explore how the ideal results from previous chapters inform our under-

standing of the more realistic world ridden with noisy measurements. In this chapter, we illustrate the working of the proposed sensor attack detection and correction methods under measurement uncertainty via a particular example of a speed measurement system.

We conclude our results in Chapter 7 and provide some possible future research directions.

1.4 Main Contributions

We summarize our main contributions to the literature as follows.

• **Extend the scope of the security index**

The notion of security index was introduced in previous literature [14, 15]. It characterises the vulnerability of a CPS against sensor only attacks for systems with zero input in a representation-free manner. We extend the concept to deal with other attack models and for more general systems with known inputs. More specifically, we introduce the following representation-free notions:

Actuator security index (Section 4.1.1);

Sensor and actuator security index (Section 5.1.1).

• **Prior attack knowledge**

In this thesis, sensor and actuator attacks detection and correction for systems with or without prior attack knowledge are considered. In general, both actuator and sensor could be attacked. However, in this work, we demonstrate that if we are given certain prior knowledge about the attack signal, for example, if we know that only sensor measurement signals are under attack (sensor only attack, Chapter 3); or only actuator input signals are under attack (actuator only attack, Chapter 4), then the corresponding detection/correction techniques will change accordingly. Furthermore, we consider the case when a specific subset of sensors/actuators is not accessible by the attacker (Section 3.3, 4.2 etc.). We then propose the corresponding techniques in order to address the problems of attack detection and attack correction. The corresponding security index is proposed in order to characterise the vulnerability of the targeted system under such prior knowl-

edge. More specifically, we introduce the following representation-free notations:

Sensor security index subject to prior sensor knowledge (Section 3.3.1);

Actuator security index subject to prior actuator knowledge (Section 4.2.1);

Sensor and actuator security index subject to prior sensor knowledge (Section 5.2.1);

Sensor and actuator security index subject to prior actuator knowledge (Section 5.3.1).

• **Computational methods for various security index**

In this thesis, we give results on how to compute the security index when different attack scenario is considered. The computational methods are discussed based on the Input/Latent/Output image representation of a system.

• **CPS attack detection and correction method**

In this thesis, attack detection and correction algorithms are proposed for different system representations and different attack models. The attack detection methods proposed in [14, 15], provide some basic understanding of the CPS security problems via kernel representation. In this thesis, we follow a similar approach. Based on the notion of the attack detectability, the proposed detection algorithms are guaranteed to achieve attack detection. With a view to address the problem of attack correction, we first propose the attack correctability for a system based on the notion of security index. We then propose attack correction algorithms. These proposed algorithms are guaranteed to achieve unique attack correction if the number of attacked sensors (actuators) is below certain upper bound. The upper bounds are stated in terms of the security index of the system.

• **Trivially secure systems**

For most of the systems considered in this thesis, we can have guaranteed attack detection/correction method when the attack signal satisfies the assumptions as mentioned in the previous dot point. However, we found that there are some systems, even if all the sensors (actuators) are attacked, it can still achieve guaranteed attack detection/correction. In this thesis, we discussed in detail the property of such systems. Detection/correction methods for trivially secure systems also proposed.

1.5 Notation Conventions

- Let $\mathbb{Z}_+ = \{0, 1, \dots\}$ and $\mathbb{R} := (-\infty, \infty)$.
- The transpose of a matrix (or signal) $\bullet$ is denoted by $\bullet^T$.
- A signal is a function of time. The N-dimensional signal $y = (y_1, ..., y_N)^T$ is denoted as $y : \mathbb{Z}_+ \to \mathbb{R}^N$.
- An $N \times N$ identity matrix is denoted by $\mathbb{I}_N$.
- The weight of a signal y, denoted by $\|y\|$, is the number of components of y that are non-zero signals.
- If $\mathcal{J}$ is a subset of $\{1, \dots, N\}$ then its complementary set is denoted by $\bar{\mathcal{J}}$, i.e., $\bar{\mathcal{J}} := \{1, \dots, N\} \backslash \mathcal{J}$.
- The l_∞ norm of a signal y, denoted by $\|y\|_\infty$, is smallest positive number M, so that $|y(t)| \leq M$ for all $t \geq 0$.
- The cardinality of a set $\mathcal{J}$ is denoted by $\|\mathcal{J}\|_c$.
- The signal that consists of the i-th components of y where $i \in \mathcal{J}$ is defined as $y^{(\mathcal{J})}$.
- The shift operator σ is defined as $\sigma y(t) := y(t + 1)$.
- Consider an $N \times M$ matrix R; then, the sub-matrix that is resulted selecting only rows with index $i \in \mathcal{J}$ is defined as $R^{(\mathcal{J},\bullet)}$. Similarly, the sub-matrix that is resulted by selecting only columns with index $i \in \mathcal{J}$ is defined as $R^{(\bullet,\mathcal{J})}$.

Chapter 2
Literature Review

2.1 Linear System Fault Detection and Isolation

2.1.1 Preliminaries

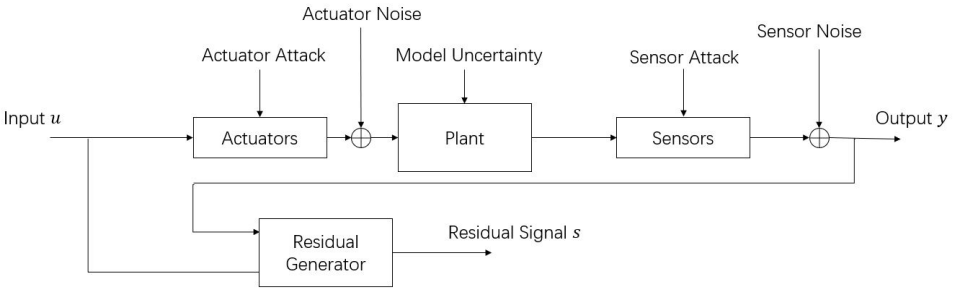

Figure 2.1: General system model

Figure 2.1 illustrates a general system model in the area of FDI. Based on the work of [34], the system model without any fault is described using the following state-space representation:

$$\begin{aligned} x(t+1) &= (A+\Delta A)x(t) + (B+\Delta B)u(t) + E_a w_a(t) \\ y(t) &= (C+\Delta C)x(t) + (D+\Delta D)u(t) + E_s w_s(t) \end{aligned} \tag{2.1}$$

where $t \in \mathbb{Z}_+$. Vector $x(t) \in \mathbb{R}^n$ is the state vector, $u(t) \in \mathbb{R}^q$ is the actuator input, $y(t) \in \mathbb{R}^m$ is the sensor measurement output. Vectors $w_a(t)$ and $w_s(t)$ are zero-mean noise or disturbance, whereas the noise is added onto the system via noise matrix E_a and E_s. The model uncertainties are captured using ΔA, ΔB, etc.

An ideal system model where we ignore any model uncertainties and any noise can be expressed in the following state space form:

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{aligned} \tag{2.2}$$

Based on the system model (2.2), we recall the following definition regarding system observability.

Definition 2.1. *(e.g., [37]) A system in the state-space representation* (2.2) *is said to be* ***observable*** *if, for any input signal u, the state signal x can be uniquely determined using only the output signal y.*

In FDI literature, a fault signal is often modelled as a additive signal on the sensor measurement output (sensor fault) or the actuator input (actuator fault). A faulty system can be described as:

$$\begin{aligned} x'(t+1) &= (A + \Delta A)x'(t) + (B + \Delta B)(u(t) + f_a(t)) + E_a w_a(t) \\ y(t) &= (C + \Delta C)x'(t) + (D + \Delta D)(u(t) + f_a(t)) + E_s w_s(t) \\ r(t) &= y(t) + f_s(t) \end{aligned} \tag{2.3}$$

where $f_a(t)$ denotes the actuator faults and $f_s(t)$ represents the sensor faults. Vector $r(t)$ represents the attacked received signal. Vector $x'(t)$ signifies the corrupted state vector. In many practical situations, fault signals $f_s(t)$ and $f_a(t)$ are state-dependent. For these cases, we can express the fault signal $f_s(t)$ (or $f_a(t)$) as a function of the corrupted state $x'(t)$. In FDI literature, it is often assumed that the input vector $u(t)$ and the corrupted output $r(t)$ are known by the user.

The notion of the residual signal is often used in FDI literature in order to reflect the inconsistency between the ideal system behaviour and the actual system behaviour. A residual signal s, as illustrated in Figure (2.1), is generated using r and u:

$$s = h(u, r). \tag{2.4}$$

In FDI literature, an output estimation process is often used in order to generate the residual signal. An output estimator generates an estimated output $\hat{y}$ using r and u:

$$\hat{y} = g(u, r), \tag{2.5}$$

followed by which, the residual signal s can be interpreted as the difference between the received output and the estimated output:

$$s = r - \hat{y}. \tag{2.6}$$

In order to achieve fault detection, the residual signal should be sensitive to the possible faults. The following concepts from [34] can be used to explain this property:

- When no fault occurs, i.e., $f_a = f_s = 0$, the residual signal s should be zero signal.
- When fault signals f_a, f_s occur, residual signal s should deviate from zero.

For fault isolation, the residual signal should be able to provide specific failure signature for different types of faults.

In the rest of this section, we review some of the methods to achieve FDI in the existing literature. This assumes significance as these methods relate to the topic of the thesis: security in CPS.

2.1.2 The Chow-Willsky Scheme

In E. Y. Chow [17], Chow-Willsky [16], FDI for an LTI system in a state-space model is achieved using a so-called parity relation based residual generator. This method is then referred to as the Chow-Willsky Scheme. If we neglect the model uncertainty, noise, and the known input signal, the system model in [16] can be expressed as follows:

$$\begin{aligned} x(t+1) &= Ax(t) + Bf_a(t) \\ y(t) &= Cx(t) \\ r(t) &= y(t) + f_s(t). \end{aligned} \tag{2.7}$$

Let us now consider a residual signal s defined as

$$s(t) = Hr(t) = HCx(t) + Hf_s(t). \tag{2.8}$$

If we select a matrix H such that $HC = 0$, then the residual signal will only correspond to $f_s(t)$; hence, it will be zero under no fault conditions, and detect a fault when a non-zero $f_s(t)$ leads to a non-zero $s(t)$.

Rather than considering a single instant in time, we may collect a history of output measurements, which leads to a residual signal that may be defined as:

$$S(t) = HR(t), \tag{2.9}$$

where

$$\begin{aligned} R(t) &= \begin{bmatrix} r(t-k) \\ r(t-k+1) \\ . \\ . \\ . \\ r(t) \end{bmatrix} \\ &= Ox(t-k) + \begin{bmatrix} 0 & 0 & . & 0 \\ CB & 0 & . & 0 \\ . & CB & . & 0 \\ . & . & . & . \\ . & . & . & . \\ CA^{k-1}B & CA^{k-2}B & . & 0 \end{bmatrix} \begin{bmatrix} f_a(t-k) \\ f_a(t-k+1) \\ . \\ . \\ . \\ f_a(t-1) \end{bmatrix} + \begin{bmatrix} f_s(t-k) \\ f_s(t-k+1) \\ . \\ . \\ . \\ f_s(t) \end{bmatrix} \end{aligned} \tag{2.10}$$

and matrix O is the observability matrix , i.e. $\begin{bmatrix} C^T & (CA)^T & ... & (CA^k)^T \end{bmatrix}^T$.

To achieve fault detection and isolation for $f_s^{(i)}$, the i-th rows of H should be able to

isolate $f_s^{(i)}$ from the state signal, which means that:

$$H^{(i)}O = 0. \tag{2.11}$$

In [16], the authors further explain that matrix H can be generated by subtracting the orthogonal projections onto O from the identity operators. More specifically, H can be chosen to consist of the independent rows of $I - O(O^TO)^{-1}O^T$. In this setup, in order to achieve fault detection, one can check that no sensor fault corresponds to $S(t) = 0$ and otherwise one decides that there exists a fault. For fault isolation, the parity vector should be able to provide specific failure signature for some specified failures. This can be achieved using the remaining freedom of $H^{(i)}$.

There are still many open research problems that for the Chow-Willsky Scheme back in the day. For example, the problem of selecting a proper parity structure to achieve fault detection and isolation is one of major concern under the Chow-Willsky Scheme, a specific parity structure can only address certain classes of faults. Then it is valid to ask, what is the design trade-off between the order (the number of states) of the system and the computational complexity? Is that possible to propose a universal parity relation that can address a border range of faults? What are the underlying assumptions for the system/fault signals regarding this statement? Later work e.g., [73] attempted to overcome certain design trade-off between the order of the system and computational complexity, [21] increase the dimension of the parity space, thus increasing the design freedom. Later on in this thesis, we will focus on proposing detection methods for any fault (attack) signals using the similar ideas of residual generation as Chow and Willsky, and we state the underlying assumptions for the proposed methods.

Another open research question followed by those works is the algorithmic guarantee. Under the Chow-Willsky Scheme, the authors did not delve into the guarantee of the proposed FDI method. For example, is that possible to propose methods that can always achieve FDI? If so, what are the underlying assumptions for the system and/or the fault signals? Against this backdrop, a guaranteed detection/correction method under certain assumption is the core of our research.

2.1.3 Unknown Input Observers

In 1982, K. Watanabe and D. H. Himmelblau introduced an FDI method based on unknown input observers (filters) [68]. The basic idea is to generate a residual signal based on state estimation error, which is decoupled from the unknown input disturbance ($E_a w_a$ from equation (2.3)). In this manner, the observer maximises the sensitivity to fault signals. Under this approach, only input faults are considered and the faulty system can be modelled as follows:

$$\begin{aligned} x(t+1) &= Ax(t) + B(u(t) + f_a(t)) + E_a w_a(t) \\ r(t) &= Cx(t). \end{aligned} \tag{2.12}$$

An unknown input observer in a state-space model is given as follows:

$$\begin{aligned} \hat{z}(t+1) &= L_1\hat{z}(t) + L_2 r(t) + L_3 u(t) \\ z(t) &:= Tx(t) \end{aligned} \tag{2.13}$$

where $\hat{z}(t)$ denotes the estimated signal of the observer. The observer matrices L_1, L_2, L_3 and T are chosen in such a manner that the estimation error of the observer $e(t) = \hat{z}(t) - z(t)$ is independent of the unknown input disturbance w_a; moreover, the error signal convergesa to zero asymptotically in the absent of fault. More specifically, it follows:

$$\begin{aligned} & e(t+1) \\ &= \hat{z}(t+1) - z(t+1) \\ &= L_1 z(t) + L_2 Cx(t) + L_3 u(t) - TAx(t) - TBu(t) - TE_a w_a(t) - TBf_a(t) \\ &= L_1 z(t) + (L_2 C - TA)x(t) - (L_3 - TB)u(t) - TE_a w_a(t) - TBf_a(t). \end{aligned} \tag{2.14}$$

Choose L_1 (with stable eigenvalues), L_2, L_3 and T, such that

$$\begin{aligned} L_2 C - TA &= -L_1 T \\ (L_3 - TB) &= 0 \\ TE_a &= 0, \end{aligned}$$

then equation (2.14) becomes

$$e(t+1) = L_1 e(t) - TBf_a(t) \tag{2.15}$$

In 1996, the work of Chen et al. [11] proposes a full order unknown input observer and gives necessary and sufficient conditions for its existence. The terminology 'full order' implies that the order of the observer is equal to the order of the system. This approach provides design freedom for the observer which can then be used to make the residual signal sensitive to fault at different locations (for example, fault at the i-th sensor). A non-linear jet engine system is subsequently used to illustrate the robust fault isolation approach.

In the 2006 work [69], the authors developed a particular system structure by regrouping the system inputs, which is suitable for unknown input observer design. In [69], a class of uncertain Lipschitz non-linear systems are considered. Based on this particular system structure, necessary and sufficient conditions for the observer's existence were provided in order to address the problem of FDI under certain assumptions.

The objective of unknown input observer-based design is to maximise the sensitivity of faults concerning a noisy system. Therefore, it is difficult to quantify the performance of such algorithms due to the presence of noise. However, if we constraint ourself to a noiseless case, it then becomes possible to quantify the performance of the FDI algorithms. Such open questions motivate the present study. In this thesis, we propose algorithms that are guaranteed to achieve attack detection/correction for noiseless systems.

In the following chapters, we will first consider a noiseless system and provide conceptual approaches to solve the problem of attack detection/correction. Then, we extend the proposed techniques into a particular engineering example where measurement noise is considered.

2.1.4 Transfer Function-Based Approach

E. Frisk [24] in [25, 26], introduced a design method based on a transfer function model to address the problem of FDI. The method aims to find all residual generators using a

numerically efficient algorithm; the method is also capable of finding residual generators with a minimum order. A minimum order residual generator means that the highest degree of the transfer function of the residual generator is at the minimum. Based on [26], 'low order systems usually imply that only a small part in the model are utilised'. Since all parts of the model are fraught with errors, this further means that few model errors will affect the residual.

As per this approach, a faulty system is modelled as follows:

$$r = G(\tau)u + W(\tau)w + L(\tau)f, \tag{2.16}$$

where w represents the noise and f represents fault signals. The transfer functions $G(\tau), W(\tau)$ and $L(\tau)$ are assumed to be proper rational transfer functions. The letter τ can be interpreted as the Laplace variable (normally denoted as s but since the letter s is used as residual signal, we changed this commonly used notation here).

A rational residual generation matrix $H(\tau)$ is designed to generate a residual signal $s = H(\tau)\begin{bmatrix} r \\ u \end{bmatrix}$. Ideally we want to have $s = 0$ for all w and u if and only if $f = 0$. To achieve that, we have

$$s = H(\tau)\begin{bmatrix} G(\tau) & W(\tau) \\ \mathbb{I} & 0 \end{bmatrix}\begin{bmatrix} u \\ w \end{bmatrix} + H(\tau)\begin{bmatrix} L(\tau) \\ 0 \end{bmatrix} f. \tag{2.17}$$

We can see that $H(\tau)$ must lie in the left null space of $M(\tau) := \begin{bmatrix} G(\tau) & W(\tau) \\ \mathbb{I} & 0 \end{bmatrix}$. Under the state-space approach, denote the system matrix $M_s(\tau)$ of the system described by (2.3) and formulate the disturbances as inputs:

$$M_s(\tau) = \begin{bmatrix} C & E_s \\ -(\tau\mathbb{I} - A) & E_a \end{bmatrix}. \tag{2.18}$$

Further, define matrix P as

$$P = \begin{bmatrix} I & -D \\ 0 & -B \end{bmatrix}. \tag{2.19}$$

Then, we have the following relationship between the matrix $M(\xi)$ and $M_s(\xi)$:

$$P \begin{bmatrix} y \\ u \end{bmatrix} = PM(\tau) \begin{bmatrix} u \\ w \end{bmatrix} = M_s(\tau) \begin{bmatrix} x \\ w \end{bmatrix} \tag{2.20}$$

In [26], it is proven that if the pair $\{A, [B\, E_a]\}$ is controllable, and $V(\tau)$ is a polynomial matrix whose rows form a minimal polynomial basis for the left null space of $M_s(\tau)$, then the rows of $W(\tau) = V(\tau)P$ form a minimal polynomial basis for the left null space of $M(\tau)$, in the sense that the number of basis required to form the left null space of $M(\tau)$ is minimum. The minimal basis is of particular interest because minimal basis implies a minimum order residual generator.

The transfer function-based residual generator design provides some intuition for our research. For instance, the system model only involves input and output signals and the state signal x is not necessarily needed while dealing with sensor/actuator faults. Thus, signal x can be treated as a latent signal. In our research, we also express our system based on an input/output model. Such a representation is then used when we are discussing attack detection. For attack correction, the auxiliary latent signal l (similar to state x) is proven to be vital and thus, we further propose an input/Latent/output model followed by attack correction algorithms.

In Frisk's works, the methods being proposed can generate a number of minimum basis observers. In other words, there exists design freedom that can be used to shape the fault-to-residual response; however, how to address such design freedom, in the sense that how to choose among those observers to achieve guaranteed fault diagnosis is not being discussed in depth. In our approach when solving attack detection and correction problems, a similar design freedom also exists. Using the notion of unimodularly equivalent system representation, we single out a class of observers that can achieve guaranteed attack detection/correction in the noiseless case.

2.1.5 Optimisation-Based Approach

The problem of FDI or robust residual generation problem can also be addressed using an optimisation-based approach. The objective for this approach is: design a mechanism that maximises the fault sensitivity while minimising the influence of the model and/or sensor measurement uncertainty.

In the work of Ding et al. [20], problems of optimising observer-based fault detection (FD) systems in the sense of increasing the robustness to the unknown disturbance and simultaneously enhancing the sensitivity to the faults are studied. The paper discusses the formulation of fault detection design problems and provides a solution for fault detection filters in the state-space model using H_∞ optimisation method.

In the work of Song and E. G. Collins Jr. [60], an H_2 estimation process is used to solve the fault detection problem for linear systems with modelling uncertainties. The H_2 estimation problem is formulated as a parameter optimisation problem in which the upper bound is minimized subject to a Riccati equation constraints. The robust H_2 estimation framework is then used in a longitudinal flight system and showed that such a framework reduces the false alarm rate compare with other methods such as a residual-based Kalman Filter estimation.

The work of D. Sauter and F. Hamelin [52] deals with the design problem of optimal filters for FDI based on the optimisation of a performance index in the frequency domain. A complete procedure for designing robust residual generators in the frequency domain is present. The design procedure is divided into two steps: the first step is to derive a specific fix-direction residual for the considered failure modes; the second step is to exploits the frequency characteristics of the system concerning the disturbance and faults. The proposed method is designed to maximise the fault to disturbance ratio.

Kalman filter approach is a special case of stochastic optimisation that utilizes linear quadratic optimisation technique. In 1971, Mehra and Peschon [40] introduced a fault detection and diagnosis technique by adopting the Kalman filtering approach. The Kalman filter is used as residual generators; the residual signal is then feed into an identifier to achieve fault detection and isolation via a series of hypothesis testing, including tests of whiteness, mean and covariance.

In many applications, the optimisation-based approach is capable to address the problem of FDI, however, there are also disadvantages for this particular approach.

One of the major concerns in this topic, or even in optimal control community is that someone needs to provide a 'criteria' to optimise. For FDI, such criteria may be, for example, optimising the 'norm' of the residual signal (such as minimising the power of the output estimation error $\|r - \hat{y}\|_{l_2}$ etc.) or optimising the 'norm' of the state estimation error. Without any doubt, such criteria are closely related to a successful FDI and in many cases have a good performance. However, those criteria are not equivalent to FDI. In this thesis, we seek for guaranteed detection/correction methods when a noiseless system is under adversarial attacks.

Another open research problem from the optimisation-based FDI is the computational complexity. The fundamental limitation for solving FDI using such an approach lies in the optimisation problem itself. Even though the system being considered in those works are LTI system at each time instant, an optimisation problem (LMI for example) still needs to be solved in real-time. Moreover, when the size of the system (the number of states, the number of inputs/outputs) becomes larger, the computational complexity will increase geometrically. In this thesis, the proposed detection/correction algorithm for noiseless systems do not involve optimisation programs. Instead, we use a bank of pre-computed observers to achieve attack correction and thus solving real-time optimisation problems are no longer needed.

2.2 Vulnerability Analysis of Cyber-Physical Systems

In 2006, the concept of Cyber-Physical System (CPS) was introduced. We can say that almost all physical infrastructure is at least monitored if not controlled using networked sensors and actuators. That is how CPSs abound. With this comes the problem of cyber security. How can we be guaranteed that the end-to-end service supported by the CPS is actually as intended? Could a cyber attack change the behaviour? How would we know?

Two critical aspects of CPSs are the main objects of interest to the attackers: sensors and actuators. Sensor outputs are the information or values produced by a system, usu-

ally monitored by an end-user. Actuator inputs drive the system to achieve specific behaviour. Typically, actuator inputs are computed based on models, and more often than not, for simplicity and robustness reasons, it will have a component based on feedback from the sensor signals. Exploiting vulnerabilities in the CPS, an attacker may change sensor and/or actuator signals in such a manner as to pursue a malicious objective, deviating from the intended CPS performance.

The other potential object of interest is the control computer, i.e., an attack on the compute engine computing the actions from the measurements; this type of attack is often referred as controller cyber attacks. In this thesis, we do not consider controller cyber attacks.

Attacks on a system may exhibit similar traits as faults and may therefore be dealt with by standard FDI techniques as in mentioned the previous section. However, it should be noted that unlike faults, attacks come from an attacker with malicious intent. Furthermore, the attacker is often assumed to have full knowledge of the targeted system, and such knowledge enables the attacker to implement attack signals that can mimic the behaviour of the system (or mimic the behaviour partially) which is difficult or impossible to detect and correct.

With this picture in mind, when dealing with attacks, a worst-case scenario thinking is always critical. For example, in FDI, different fault models can potentially provide extra parity relation freedom; some of the fault-tolerant filters assume a priori knowledge (statistical or temporal) of the fault signal [8]. While in a CPS attack, such assumptions and design freedoms are off the table. In the CPS community, when the adversary attacks a sensors and/or actuator it can change the signal to any arbitrary values without any restriction (statistical, shape or otherwise).

The only limitation that we impose on the attack signal is the total number of sensors and/or actuators being attacked. In most of the CPS security literature [12,14,15,22,41,44] etc., it is assumed that an upper bound on the number of attacked sensor and/or actuator is given, but the user does not know which ones are being attacked. In our research, we also make this assumption.

In this section, we review the literature around characterizing the vulnerability of a

system against such attacks.

Analogous to the fault detection and isolation literature, a general system and attack model can be expressed in the following form:

$$\begin{aligned} x(t+1) &= Ax(t) + B(u(t) + \eta_a(t)) + E_a w_a(t) \\ y(t) &= Cx(t) + D(u(t) + \eta_a(t)) + E_s w_s(t) \\ r(t) &= y(t) + D_s \eta_s(t) \end{aligned} \tag{2.21}$$

where $\eta_a(t) \in \mathbb{R}^q$ represents the additive actuator attack signal; $y(t) \in \mathbb{R}^m$ denotes the system output without the additive sensor attack $\eta_s(t)$. Signal $r(t) \in \mathbb{R}^m$ is the attacked received signal. It is usually assumed that the input signal $u(t)$ and the received signal $r(t)$ are known and measured signal, i.e., in the present, their past is known.

Measurement redundancy is an essential concept for CPS security. In most of the existing CPS security literature, the system is assumed to have a certain level of measurement redundancy. In state-space representation, such assumptions could be that the number of sensors is larger than the state dimension, or the output matrix C has full row rank etc. However, as a starting point in our work, we do not make such assumptions, since our research aims to propose a general framework that can be applied to any LTI systems. The only assumption we make in terms of a state representation is that we assume that the system is observable as in Definition 2.1.

2.2.1 Security Index $\delta(\Sigma)$

In 2016, the works of Chong and Kuijper [14,15] proposed the concept of 'security index' as a quantitative representation-free system parameter that expresses the vulnerability of a discrete-time LTI system with respect to sensor attacks.

In the previous Section 2.1, methods and concepts to address FDI problems depend on the model and its representation. How to characterise the system vulnerability in a representation-free manner was not discussed in the existing FDI literature. In [14, 15], the authors provide their novel approach regarding a representation-free concept that characterises the system vulnerability when a system is under adversarial attacks. A

representation-free approach is more general when compared with previous representation-dependent approaches, in the sense that such a representation-free system parameter can be applied to various system representations, such as state-space, polynomial, transfer function-based representations etc. In this thesis, we follow such a representation-free approach as in [14, 15].

In these two works, Chong and Kuijper consider an ideal system where they ignore the model and measurement uncertainty, focusing purely on sensor attacks. For consistency purposes, the system model in [14, 15] in a state-space representation is described as follows, but do note that [14, 15] also looks at various other system representations:

$$\begin{aligned} x(t+1) &= Ax(t) \\ y(t) &= Cx(t) \end{aligned} \tag{2.22}$$

$$r(t) = y(t) + \eta_s(t). \tag{2.23}$$

The behaviour $\mathcal{B}$ of the system is defined as the set of all possible output signals (trajectories) that satisfies the attack-free system dynamics (2.22). The authors of [14, 15] assumed that a restricted number of sensors are under attack. This assumption can be captured using the following model:

$$\begin{aligned} \eta^{(i)} &\neq 0 \text{ for } i \in \mathcal{J} \subseteq \{1, ..., m\} \\ \eta^{(i)} &= 0 \text{ for } i \notin \mathcal{J}, \end{aligned} \tag{2.24}$$

where m is the total number of outputs, $\mathcal{J}$ denotes the set of attacked sensors and $\|\mathcal{J}\|_c$ is upper bounded by certain value. Detectability and correctability of a system are defined in terms of the system behaviour $\mathcal{B}$, where $\mathcal{B}$ is the set that contains all the possible attack-free output signals y. The security index of the system Σ is defined in a representation-free manner as follows:

$$\delta(\Sigma) := \min_{0 \neq y \in \mathcal{B}} \|y\|, \tag{2.25}$$

where $\|y\|$ is the weight of signal y. From the above equation, we can conclude that the security index is defined as the minimum weight among all possible output signals.

From another point of view, it means that we are searching for the sparsest output trajectory inside the behaviour of the system.

It has been proven in [14,15] that if the number of sensors being attacked $\|\eta\| < \delta(\Sigma)$, then attack detection is possible. On the other hand, if the attacker is capable of attacking more than (or equal to) $\delta(\Sigma)$ sensors, then it is possible for the attacker to mimic the (attack-free) behaviour, thus being undetectable.

It has also been proven in [14, 15] that if the number of sensors being attacked $\|\eta\| < \delta(\Sigma)/2$, then a unique attack correction is possible. In other words, in order to implement an undetectable (uncorrectable) attack, the attacker must be able to attack at least $\delta(\Sigma)$ ($\delta(\Sigma)/2$) sensors.

Finding the security index of system Σ from equation (2.25) is a difficult problem since it requires a search over the entire system behaviour. In [14, 15], four different methods of computing the security index are presented.

- Via state-space representation:

$$\delta(\Sigma) = \min_{j\in\{1,\dots,m\}} \min_{x\in V_j} |supp(Cx)|, \tag{2.26}$$

  where V_i is the subspace spanned by all eigenvectors corresponds to eigenvalue λ_i of matrix A, and $supp(\bullet)$ denotes the support of a vector space.

- Via coding matrix O': the coding matrix of system Σ is defined as

$$O' = \begin{bmatrix} O_1 \\ O_2 \\ \vdots \\ O_m \end{bmatrix}, \tag{2.27}$$

  where matrix O_i denotes the observability matrix with respect to the i-th sensor output, i.e., $O_i = \begin{bmatrix} C_i^T & (C_iA)^T & \dots & (C_iA^k)^T \end{bmatrix}^T$. Then the security index $\delta(\Sigma) = m - l$ where l is the largest integer for which there exists a subset $\mathcal{J} \subseteq \{1, \dots, m\}$ with $\|\mathcal{J}\|_c = l$ such that the kernel of $O'_{\mathcal{J}}$ is not zero.

- Via a check matrix H: a non-zero check matrix H is defined as

$$H = \begin{bmatrix} H_1 & H_2 & \dots & H_m \end{bmatrix} \text{ such that } HO' = 0. \tag{2.28}$$

  Then the security index $\delta(\Sigma) = \text{spark}(H)$ where $\text{spark}(H)$ is defined as the smallest integer l for which there exists a subset $\mathcal{J} \subseteq \{1, ..., m\}$ with $\|\mathcal{J}\|_c = l$ such that the kernel of $H_{\mathcal{J}}$ is not zero.

- Via a kernel representation: apart from the state space representation, a kernel representation can also be used for describing a system's dynamics. In [14, 15], the kernel representation of the system Σ is given by

$$R(\sigma)y = 0 \tag{2.29}$$

  where $R(\xi)$ denotes an $m \times m$ polynomial with non-zero determinant. The security index $\delta(\Sigma)$ is then given by the smallest integer l for which there exists a subset $\mathcal{J}$ with $\|\mathcal{J}\|_c = l$ such that $R^{(\mathcal{J},\bullet)}(\xi)$ is not left unimodular.

The kernel representation of the system also serves as the starting point of our research. The advantage of a kernel representation is that it is a minimalist way of expressing a system's dynamics, in the sense that the only focus for the kernel representation is the input and output signals. As stated before, the kernel representation is very general in the sense that under the observability assumption as in Definition (2.2), the kernel representation can be transferred to other representations (state-space for example), in a sense that all the different representations state the same inputs/outputs relationship. Together with a behavioural approach, it enables us to present condensed statements and theorems regarding CPS security.

The concept of security index characterises the vulnerability of an LTI system against sensor attack. The representation-free property of the security index indicates that the notion of security index is a true system property. This allows us to use such a notion for different system representations. Our research and this thesis are discussed based on this concept.

There are several open research problems followed by [14,15], for example, the system model only involves output signals. In this thesis, we extend the scope to input/output systems. We also note that the concept of the security index is not complete. When the behaviour is $\{0\}$ only, the security index is undefined. This may not be a major problem without considering inputs. However for systems with inputs, more attention is needed. Later on in this thesis, we define this class of system as 'trivially secure' systems.

Another open research question is actuator attack. The security index in [14, 15] only addresses sensor attacks, and thus it is valid to ask is it possible to extend this concept to other attack scenarios, such as actuator only attack; actuator and sensor attack etc. In this thesis, we extend the concept of security index to address those different attack scenarios.

Last but not the least, an attack correction algorithm is missing in [14, 15]. In this thesis, the problem of attack correction is being addressed.

2.2.2 Security Indices α_i

In [50] [30], the terminology 'security indices' was used as indices to describe the vulnerability of a power system in steady state. Later on in 2016, Sandberg and Teixeira [51] generalised this concept to a class of dynamical systems. The 'security index α_i for sensor i' in [51] expresses the vulnerability of sensor i with respect to attacks. More specifically, α_i can be interpreted as follows: if an attacker wants to attack the i-th sensor of a system, what is the minimum number of other sensors the attacker needs to attack in order to be undetectable?

In [51], the authors extend the concept of security indices for a dynamical system with zero input ($B, D = 0$) under noisy environment. The system and attack model being considered in [51] are as follows:

$$\begin{aligned} x(t+1) &= Ax(t) + E_a w(t) + B_a \eta(t) \\ y(t) &= Cx(t) + E_s w(t) \\ r(t) &= y(t) + D_s \eta(t), \end{aligned} \tag{2.30}$$

where $w(t) \in \mathbb{R}^l$ denotes the unknown disturbance signals, the disturbances influence

the system via known matrices E_a and E_s as in Section 2.1. Vector $\eta(t) \in \mathbb{R}^q$ represents the adversarial attack signal at time t, the attack signal influences the system via B_a (actuator attack matrix) and D_s (sensor attack matrix). In the presence of noise w, the objective in [51] is to ascertain whether the existence of a disturbance will 'mask' the attack. If this is the case, the operator is not able to distinguish between attacks and disturbances, and thus, cannot conclude whether an attack is present or not. To address that issue, the authors assume distinct measurements, namely:

$$\text{rank} \begin{bmatrix} E_a \\ E_s \end{bmatrix} = l, \quad \text{rank} \begin{bmatrix} B_a \\ D_s \end{bmatrix} = q, \quad \text{rank} \begin{bmatrix} C \end{bmatrix} = m, \tag{2.31}$$

where m is the number of outputs.

In [51], the authors further assume that the attack signal follows a specific pattern: $\eta(t) = \eta_0 z_0^k$, $\eta_0 \in \mathbb{C}^q$, $z_0 \in \mathbb{C}$. In [51], an attack signal is **undetectable** if there exists a simultaneous disturbance signal $w(t)$ and an initial state $x(0)$ such that $y = 0$. Minimum resources needed by the attacker to achieve undetectability are discussed in [51].

An attack signal is undetectable if there exists a x_0 and w_0 such that

$$\begin{bmatrix} A - z_0 I & E_a & B_a \\ C & E_s & D_s \end{bmatrix} \begin{bmatrix} x_0 \\ w_0 \\ \eta_0 \end{bmatrix} = 0 \tag{2.32}$$

If an attacker would like to target the i-th element, i.e., $\eta^{(i)} \neq 0$ and remain undetected, this may require the attacker to target several other elements. For each i, a security index α_i is defined as the minimum number of elements (other than the i-th element) the attacker needs to attack in order to be undetectable:

$$\begin{aligned} \alpha_i := & \min_{|z_0| \geq 1, x_0, w_0, \eta_0^{(i)} \neq 0} \|\eta_0^{(\bar{i})}\| \\ \text{subject to } & \begin{bmatrix} A - z_0 I & E_a & B_a \\ C & E_s & D_s \end{bmatrix} \begin{bmatrix} x_0 \\ w_0 \\ \eta_0 \end{bmatrix} = 0. \end{aligned} \tag{2.33}$$

Then, for any attack position i, an undetectable attack η exists if and only if $\|\eta^{(\bar{i})}\| > \alpha_i$; attack correction (identification) is possible if and only if $\|\eta^{(\bar{i})}\| < \alpha_i/2$.

The concept of security indices α_i shows similarities with security index $\delta(\Sigma)$ in some sense. Both these concepts describes the vulnerability of a system with respect to sensor only attack. The notion of α_i is related to the vulnerability for systems with disturbances, while the notion of $\delta(\Sigma)$ is related to noiseless systems. The notion of α_i can be interpreted as specific prior sensor knowledge, namely, sensor i is assumed to be attacked. In later chapters in our thesis, we also take into account certain prior knowledge, albeit from a different perspective, namely we are given a subset of sensors and/or actuators that is guaranteed to be attack-free.

The works around security indices [50] [30] and [51] also left some open research problems. One of the open research problems followed by these works is the attack model. In those works, the attack signal is assumed to follow a specific geometric model. In this thesis, we make no such assumptions with respect to the attack pattern.

2.2.3 Strong Observability

In 2015, the work of [12] studied the CPS security subject to dynamical sensor attacks and related such security issue to the strong observability of a system. The authors of [57] discussed the problem of state estimation subject to sensor and/or actuator attack by introducing the notion of 'sparse strong observability'. In both works, the vulnerability of a system under adversarial attacks are being discussed.

In [12], the system Σ and sensor attack model satisfies the following equations:

$$\begin{aligned} x(t+1) &= Ax(t) \\ y(t) &= Cx(t) \\ r(t) &= y(t) + D_s\eta_s(t) \end{aligned} \tag{2.34}$$

The strong observability of system Σ is discussed in the following way:

Recall that the observability matrix $O_k = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^k \end{bmatrix}$ as in Section 2.1.2, the input-unobservable subspace over k steps is called $\mathcal{L}_k$. Then there exists an attack sequence $\mathcal{H}_s(k-1) \in \mathcal{L}_k$ with $\mathcal{H}_s(k-1) = \begin{bmatrix} \eta_s(0)^T & \eta_s(1)^T & ... & \eta_s(k-1)^T \end{bmatrix}^T$ that satisfies

$$O_{k-1}x + (I_k \otimes D_K)\mathcal{H}(k-1) = 0 \tag{2.35}$$

where D_K is a submatrix of D which describes the 'attack mode' of the attack signal. The subscript K indicates the attack position where K represents the set that corresponds to the indices of the non-zero attacks; $\otimes$ is the Kronecker product, i.e., component-wise multiplication.

The weakly unobservable subspace of a system Σ, denoted as $\mathcal{V}(\Sigma)$, is defined as the system's input unobservable subspace over n steps $\mathcal{L}_n$ where n in this section represents the number of states. In [12], a system Σ is called **strongly observable** if its weakly unobservable subspace is trivial, i.e., $\mathcal{V}(\Sigma) = 0$.

In [57], input/output systems are considered. An equivalent definition **strong observability** is stated as follows: a system is strongly observable if the state signal x and input signal u can be uniquely determined using output signal y.

It is also shown in [12] that an undetectable sensor only attack exists if and only if the system is not strongly observable. Equivalently, an undetectable attack exists if and only if there is an eigenvector v of A such that vector Cv lies in the range space of D_K.

The later work [57] proposes a state estimation method for a noiseless system ($E_a, E_s = 0$) in the presence of both actuator and sensor attacks. State signal plays a vital role not only in [57], but also for general CPS security literature, especially when actuator faults/attacks are present. An attack on the actuator can lead to a corrupted state which is generally not being measured directly by the end-user. A successful estimation of the state signal can provide more insights regarding the attacked system.

The detectability and correctability of such systems is discussed using the notion of sparse strong observability. In [57], the system model is:

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) + B\eta_a(t) \\ y(t) &= Cx(t) + D(u(t) + \eta_a(t)) \\ r(t) &= y(t) + \eta_s(t). \end{aligned} \tag{2.36}$$

Consider the observability matrix O and the invertibility matrix defined as follows:

$$\mathcal{N}_{(A,B,C,D)} := \begin{bmatrix} D & 0 & . & 0 \\ CB & D & . & 0 \\ . & CB & . & 0 \\ . & . & . & . \\ . & . & . & . \\ CA^{n-2}B & CA^{n-3}B & . & D \end{bmatrix}.$$

The work [57] deals with subsets of sensors and actuators; thus for a subset of sensors $\mathcal{J} \subseteq \{1, ..., m\}$ and for a subset of actuators $\mathcal{K} \subseteq \{1, ..., q\}$, we refer to $O_{\mathcal{J}}$ as the observability matrix relates to sensor subset $\mathcal{J}$ and we denote $\mathcal{N}_{\mathcal{K}\to\mathcal{J}}$ as the invertibility matrix relates to actuator subset $\mathcal{K}$ and sensor subset $\mathcal{J}$.

In order to characterise the vulnerability of a system against adversarial attacks, the notion of (j, k) **- sparse strong observability** is proposed: a system is (j, k) - sparse strong observable if for any $\mathcal{J} \subseteq \{1, ..., m\}$ with $\|\mathcal{J}\|_c \leq j$ and $\mathcal{K} \subseteq \{1, ..., q\}$ with $\|\mathcal{K}\|_c \leq k$, the system $(A, B^{(\bullet,\mathcal{K})}, C^{(\bar{\mathcal{J}},\bullet)}, D^{(\bar{\mathcal{J}},\mathcal{K})})$ is strongly observable.

An upper bound for the number of sensors/actuators being attacked before a successful state reconstruction is also presented in [57]. As a result, given that a system is $(2j, 2k)$- sparse strongly observable, state reconstruction is possible under the assumption that the number of attacked sensors and actuators are bounded by j and k, respectively. Given that the state can be reconstructed, and the inputs are known, the attack signal can be reconstructed.

This modified notion of strong observability is the key for formalising redundancy across sensors. The underlying intuition for such sparse-strong observability is that the states and inputs can be uniquely determined even if some of the sensor measurements

are removed.

The notion of strong observability and sparse strong observability are powerful tools when solving CPS attack detection and correction problems, especially when dealing with actuator attacks. Recall that for sensor only attacks, observability is a standard assumption, in the sense that if the system is not observable, then the problem of sensor attack correction is hard to solve. Since reconstructing the correct state signal plays a vital role in attack correction. When the system is not observable, even with no attacks, reconstructing the state signal is challenging. In terms of correcting sensor and actuator attacks, we are aiming to correct all actuator attacks and some (but not all) sensor attacks. For this reasoning, the notion of strong observability serves a similar purpose for actuator and attacks. In later chapters, when we present our methods for solving sensor and actuator attacks, we also make such strong observability assumption.

One of the research problems that stem from [12] and [57] is to describe the system vulnerability in a representation-free manner. In both [12, 57], the analysis is based on a system representation (state-space). In our thesis, we extend the representation free concept of the security index in [14, 15] to actuator (and sensor) attacks.

Another open research problem is that in [57], the condition for a successful attack correction is stated using two parameters: the number of sensor attacks and the number of actuator attacks. In this thesis, instead of assuming sparse strong observability, we assume that the system is strongly observable. Under this assumption, we then present our notion of the correction index, which characterises the correctability of a system under sensor and actuator attacks. We also note that under the strong observability assumption, it is possible to propose a guaranteed attack correction algorithm even when all the actuator are corrupted.

2.3 Attack Detection and Correction Methods

In order to address the potential damage and economic loss caused by the attacker, we need to propose some counter-measurement methods to prevent the consequences of an attack, namely, attack detection and attack correction methods. The purpose of attack

detection is to detect the presence of attack signals. Thus, an alarm signal can be triggered to warn the end-user that an attack has occurred. The purpose of attack correction is to enable an algorithm to achieve automatic attack correction so that the impact of such attacks can be reduced. In this section, we review some methods of implementing attack detection and correction in the existing literature.

2.3.1 Sensor Attack Detection and Correction

Sensor attack security (detection and correction) is often related to the concept of state estimation/state reconstruction (see e.g., [41,44]). Since sensor attacks influence the measurement outputs while the state signal remains the actual state, if one can reconstruct the state signal correctly, then it is trivial to generate the attack-free sensor outputs. In this section, we review two different approaches for solving the problem of sensor attacks.

Relaxed l_0 optimisation method

In [22], a state estimation method was used to solve the problem of sensor attack. The problem set up is the same as (2.22). An ideal state estimator (or decoder) has the following form:

$$\min_{\hat{x}\in\mathbb{R}^n} \|r - C\hat{x}\|, \tag{2.37}$$

where $\hat{x}$ is the estimated state signal, $\| \bullet \|$ represents the l_0 norm of a signal. The problem of the ideal state estimator is that since l_0 norm is highly non-convex. In fact this optimisation problem is known to be NP-hard (see e.g., [65]). The authors of [22] proposed a relaxation program to approximate such l_0 optimisation. Moreover, the relaxed optimisation program takes into account a specific segment of the outputs in time. The length of the segment T is proportional to the number of states of the system:

Define the linear map $\Phi^T : \mathbb{R}^n \to \mathbb{R}^{m\times T}$

$$x(t) \to \begin{bmatrix} Cx(t) & CAx(t) & ... & CA^{T-1}x(t) \end{bmatrix}.$$

Furthermore, denotes $Y_T(t)$ as the $m \times T$ matrix formed by concatenating the output measurements $y(t)$ as follows:

$$Y_T(t) = \begin{bmatrix} y(t) & y(t+1) & ... & y(t+T-1) \end{bmatrix} \in \mathbb{R}^{m \times T}$$

Define the l_0 norm of a matrix $M \in \mathbb{R}^{m \times T}$ with rows $M_1, ..., M_m \in \mathbb{R}^T$ as the number of non-zero rows in M, i.e.,

$$\|M\| = \|i \in \{1, ..., m\} | M_i \neq 0\|_c.$$

Note that under such notation, the ideal state estimator can be written as:

$$\underset{\hat{x}(t) \in \mathbb{R}^n}{\operatorname{argmin}} \|Y_T(t) - \Phi^T \hat{x}(t)\|.$$

Analogously, the authors define an l_1 state estimator as follows:

$$\underset{\hat{x}(t) \in \mathbb{R}^n}{\operatorname{argmin}} \|Y_T(t) - \Phi^T \hat{x}(t)\|_{1/r},$$

whereby definition, $\| \bullet \|_{1/r}$ is the sum of the l_r norm of the rows of the matrix M. Accordingly, the non-convex l_0 norm can be relaxed into a convex optimisation problem. Typically, the l_r norm is chosen to be l_1 or l_2.

However, one concern that was not discussed in [22] is the performance of the proposed method. Since in the work of [9], the l_0 program is equivalent to l_1/l_r norm under specific assumptions: matrices A and C need to be Gaussian random matrices, the number of sensors being attacked should not exceed $0.914\sqrt{m/2}$, the attack signal should also be Gaussian etc. For this reason, since the example proposed by [22] exams an actual system that is not randomly generated, the state estimation method has no correction guarantee even when the number of attacked sensors is below the proposed upper bound. This can be seen in the numerical example section of [22].

Moreover, the problem of attack detection is not being addressed as a separate topic in [22]. In fact, in subsequent chapters, we will demonstrate that a successful attack

detection is more straightforward compared with attack correction.

SMT-based observer design

The authors of [56] proposed a novel multi-modal Luenberger (MML) observer based on efficient Satisfiability Modulo Theory (SMT). They consider a MIMO system under noisy environment as follows:

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) + w_a(t) \\ y(t) &= Cx(t) + w_s(t) \\ r(t) &= y(t) + \eta_s(t) \end{aligned} \tag{2.38}$$

The objective in [56] is to find a secure Luenberger Observer with estimated state signal $\hat{x}(t)$ such that

$$\limsup_{t\to\infty} \|x(t) - \hat{x}(t)\|_2 \leq \rho(\max w_a, \max w_s), \tag{2.39}$$

where the upper bound $\rho()$ is a function of the bounds on the noise.

The state observer is discussed under the scope of the s-sparse observable system. Meaning the system pair $(A, C^{(\mathcal{J},\bullet)})$ with $\|\mathcal{J}\|_c = m - s$ is observable. Based on such notion, it is proven that there exists an observer which meets the requirement (2.39) if and only if for every subset $\mathcal{J}$ with at least $m - 2s$ elements, the pair $(A, C^{(\mathcal{J},\bullet)})$ is observable. Under such an assumption, the observer has the following state-space representation:

$$\begin{aligned} \hat{x}_{\mathcal{J}}(t+1) &= A\hat{x}_{\mathcal{J}}(t) + Bu(t) + L_{\mathcal{J}}(y_{\mathcal{J}}(t) - \hat{y}_{\mathcal{J}}(t)) \\ \hat{y}_{\mathcal{J}}(t) &= C^{(\mathcal{J},\bullet)}\hat{x}_{\mathcal{J}}(t), \end{aligned} \tag{2.40}$$

where $L_{\mathcal{J}}$ can be chosen such that the eigenvalues of $A - L_{\mathcal{J}}C^{(\mathcal{J},\bullet)}$ are strictly within the unit disk since the pair $(A, C^{(\mathcal{J},\bullet)})$ is observable. In the presence of noise, the aim of this work is to select an estimated state whose estimation error is no worse than the one generated by the attack-free sensors. If the noise is being neglected, then the estimation

error should be zero.

The authors in [56] realised that finding such observer using traditional method (brute force, for example) is high in memory complexity and computational complexity. For this reason, the work [56] also focuses on the development of a scalable observer architecture to overcome such disadvantage.

The proposed SMT-based program for finding the optimum observer involves three objectives: (i) hypothesize which sensors are attack-free and hence, select the mode of the MML-observer; (ii) check whether the selected set of sensors is indeed attack-free; and (iii) generate conflicts to indicate the sensor selection for next iteration in order to accelerate the search.

The basic idea of such program is as follows: first consider the l_2 norm of the residual signal as defined similarly in (2.6), the residual value is expressed as follows:

$$\|Y - O_{\mathcal{J}}\hat{x}_{\mathcal{J}}\|_2^2,$$

where $O_{\mathcal{J}}$ is the observability matrix for pair $(A, C^{(\mathcal{J},\bullet)})$. If the value is smaller than certain upper bound (proportional to the largest eigenvalue of the A matrix), we can then accept the state estimate based on this sensor selection. Otherwise, update set $\mathcal{J}$ based on the previous selected sensor set.

The convergence of the algorithm is proven, and runtime performance is being analysed. The focus and advantage of the proposed algorithm lies in its computational efficiency. It is shown that the proposed algorithm scales well for large systems (up to 5000 sensors). However, the performance in terms of the estimation error rate is not discussed in detail in [56].

In fact, the residual signal $Y - O_{\mathcal{J}}\hat{x}_{\mathcal{J}}$ can be used for attack detection. In their subsequent work [42], an attack detection Algorithm 1 was proposed based on a Kalman filter and such a residual signal.

The attack detection algorithm in [42] achieves attack detection for a chosen set $\mathcal{J}$ based on the block residual r_b.

A major comment for this detection algorithm is that such algorithm can only detect whether an attack has occurred in the particular set $\mathcal{J}$ with $\|\mathcal{J}\|_c = m - 2s$, while how to

Algorithm 1 Attack detection method in [42]

1: Run a Kalman filter that uses all measurements from sensors indexed by $\mathcal{J}$ with $\|\mathcal{J}\|_c = m - 2s$ until time $t_1 - 1$ and compute the estimated $\hat{x}_{\mathcal{J}}(t_1)$.
2: Recursively repeat the previous step $N - 1$ times to estimate a state signal $\hat{x}_{\mathcal{J}}$ for a time window from $t = t_1$ to $t = t_1 + N - 1$.
3: Calculate the block residual signal γ_b for a time window from $t = t_1$ to $t = t_1 + N - 1$:

$$\gamma_b = Y - O_{\mathcal{J}}\hat{x}_{\mathcal{J}},$$

4: **if** the block residual signal γ_b passes the test below: **then** decide that there is no attack for those chosen set $\mathcal{J}$ at time t_1.

$$\mathbb{E}(\gamma_b\gamma_b^T) - (OP_{\mathcal{J}}^*O^T + M_{\mathcal{J}}) \leq \epsilon(\lambda_{min}, \mathcal{J}, n),$$

where $\mathbb{E}$ represents the sample average using N-sample values, $P_{\mathcal{J}}^*$ is the error covariance matrix of the state signal corresponds to set $\mathcal{J}$. The threshold function ϵ is a function of the minimum eigenvalue λ_{min}, for the chosen set $\mathcal{J}$.
5: **else** decide attack occurred for the chosen set $\mathcal{J}$ at time t_1.
6: **end if**

directly perform attack detection for the overall system remains an open question. In our research and later on in this thesis, we propose attack detection and correction algorithms that deal with the entire system.

2.3.2 Actuator (and Sensor) Attack Detection and Correction

As mentioned before, the actuator is another critical feature of a CPS which is also vulnerable against attacks. In general, actuator attack is not a dual problem of the sensor attack. This means that some of the methods that can be used to solve sensor attack, such as state estimation methods, may not be applicable for the actuator attack case. This is because the actuator attacks influence the state signal directly, and thus new methodologies are needed to solve the problem of actuator attack. The actuator attack detection/correction is still an open research area, and there are few contributions in the literature that address the actuator only case. In this section, we review some of the existing literature around actuator (and sensor) attack detection/correction.

SMT-based approach

The 2017 authors of [57] extended the system set up and attack model as seen in [56] to a system under both actuator and sensor attacks, using the notion of strong observability. The objective of this work is to reconstruct the state signal $x(t)$ as in (2.36), given the received signal r and the input signal u. The system model, attack model and the corresponding assumptions are discussed in Section 2.2.3. In this section, we review the algorithms for the purpose of state reconstruction in [57].

The algorithm consists of two blocks that interact with each other: An SAT solver and a Theory solver. The SAT solver is used to update the sensor/actuator selection while the Theory solver solves the initial state based on the SAT solver selection. Given the system is $(2j, 2k)$-sparse strongly observable, the number of attacked sensors and actuators are assumed to be bounded by j and k. Initially, define a set Φ_B of the attacked actuators and sensors constrained on the upper bound (j, k), after which Algorithm 2 is proposed for state reconstruction.

Algorithm 2 State reconstruction method in [57]

1: **procedure** (A, B, C, D, r, u, j, k)
▷ Given system matrices (A, B, C, D) input u, output r, upper bound j, k, compute state signal x and attack set.
2: status $\leftarrow$ UNSAT.
3: $\Phi_{cert} \leftarrow$ True.
4: $\Phi_B \leftarrow (|\eta_s| \leq j) \bigwedge (\|\eta_a\| \leq k)$.
5: **while** status == UNSAT **do**
$\Phi_B \leftarrow \Phi_B \bigwedge \Phi_{cert}$;
$(\eta_s, \eta_a) \leftarrow$SAT-solver$(\Phi_B)$
(status, x) $\leftarrow$Theory-solver.check(supp$(\bar{\eta}_s)$, supp(η_a))
$\Phi_{cert} \leftarrow$Theory-solver.certificate(supp$(\bar{\eta}_s)$, supp(η_a))
6: **end while**
return (x, η_s, η_a)
7: **end procedure**

The Theory-solver.check uses an optimisation program, returns an optimum state and input signal based on the received signal r. Substitute the optimum (x, U) into the system dynamics and check whether such optimum solution corresponds to the received signal r, before denoting the status as SAT. The Theory-solver.certificate function provides a

reason (e.g., which sensor/actuator cause the UNSAT), then update Φ_B.

In [57], the Theory-solver.certificate finds the attacked sensors/actuators iteratively. Two different approaches for the Theory-solver.certificate were designed for shortening the running time and to improving the performance of the program. The example shows that for a system with 20 sensors and 10 actuators, the run time of this program is in the order of 100-1000 second.

However, as being mentioned before, new approaches that bypass such an optimisation problem and in effect, further improving the correction performance is still an open research problem. In a later chapter in this thesis, we provide our novel approaches on solving the detection/correction problems using pre-computed observers rather than optimisation approach.

Stabilisation for State-Dependent Sensor and Actuator Attacks

In 2017, the work [36] proposed a novel adaptive control architecture to address the security and safety issues in CPS. The authors developed an adaptive controller that guarantees uniform ultimate boundedness of the continues-time closed-loop dynamical system when facing adversarial sensor and actuator attacks. In [36], it is assumed that the sensors provide direct measurements of the states (no C, D matrices in state-space) and that the attack signals on both sensor and actuator are state-dependent.

The system model is as follows:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) \\ x'(t) &= x(t) + \eta_s(t, x(t)) \\ u'(t) &= u(t) + \eta_a(t, x(t)), \end{aligned} \tag{2.41}$$

where the first equation represents the attack-free continues time system model. Signal x' and u' is the attacked state signal (in [36], state = output) and attacked actuator signal. Signal $\eta_s()$ and $\eta_a()$ denote the state-dependent sensor attacks and actuator attacks. The

objective of this work is to design a controller of the following form:

$$u(t) = Kx'(t) + v(t), \tag{2.42}$$

where $v(t)$ denotes a corrective signal that suppresses the effect of the state-dependent attacks.

Further assume that the actuator attack signal can be parameterized as $\eta_a(t, x(t)) = W^T(t)\phi(x(t))$ where $W(t)$ is a waiting matrix and $\phi()$ denotes a non-linear function with a known structure. Since the state signal $x(t)$ is unknown, we rewrite the system representation as follows:

$$x'(t) = Ax(t) + B[u(t) + W^T(t)\phi(x'(t)) + \rho(t, x(t))],$$

where $\rho()$ is an unknown but bounded signal. Under such a system setup, one can use the corrective signal given by the following equation:

$$v(t) = -\hat{\mu}(t)Kx'(t) - \hat{W}^T(t)\phi(x'(t) - \hat{\rho}(t)\text{sgn}(B^T P x'(t))),$$

where $\hat{\mu}(t)$ denotes a scalar estimation depends on the sensor uncertainty, $\hat{W}(t)$ is the estimation of the parametric uncertainty; $\hat{\rho}$ is the estimation of the unknown bound. The rest of this work discusses how to choose those three variables which then can be used for the closed-loop control system.

Two different dynamical controllers were proposed in [36] for solving time-varying state-dependent sensor and actuator attacks. The first attempt to design a controller that is uniformly bounded for all initial state x_0, sensor uncertainty $\hat{\mu}(t)$, parametric uncertainty $\hat{W}(t)$ and unknown parameter $\hat{\rho}$. However, such controller architecture is discontinued because of the signum function sgn(). Such discontinuity can lead to a chattering phenomenon, which is undesirable. In order to lower the impact of such chattering, a smoothing function is used, that is, by replacing the signum function sgn() by the hyperbolic tangent function tanh(), which resolves the chattering phenomenon without compromising the performance of the controller. The second attempt target to design a

controller that is Lyapunov stable for all four parameters $(x_0, \hat{\mu}(t), \hat{W}(t), \hat{\rho})$. It is proven that the second design achieves Lyapunov stability. Finally, in [36], the authors provide a numerical example to illustrate the efficiency of the proposed control architecture.

The actuator attack model in [36] is assumed to be state-dependent. In general, the attack signals (both actuator and sensor attacks) are not necessarily state dependent, especially when we are dealing with malicious attacks, since a smart attack signal can be arbitrary. For this reason, our work in the following chapters do not make such an assumption.

2.4 Open Research Questions

CPS security is a relatively new research area. From the exploration of the existing literature, it can be seen that there are still many open research problems in this area.

- As in the classic FDI literature, the CPS security against attacks can be divided into two sub-problems: attack detection and attack correction. The problem of attack detection has received less attention in the existing literature. In fact, if the system model is described properly, attack detection and attack correction can be solved separately. Moreover, generally speaking, the problem of attack detection is much simpler as compared to attack correction, and in many applications, attack detection itself is a satisfactory outcome. Thus, we address the problem of attack detection as a separate topic in this thesis.

- In some of the literature on CPS security, attack correction methods are proposed based on an optimisation-based approach. Such an approach requires to solve one or more optimisation problem at each time instance, taking into account a pack of historical input/output data. For large scale systems, this often implies high computational complexity. How to by-pass such an approach is still under development. In this thesis, we provide novel approaches to address the problem of attack correction using a bank of pre-computed observers that do not involve any optimisation process.

- How to perform attack detection and attack correction based on specific sensor and/or actuator prior knowledge is still under development. In this thesis, we address this topic to some extent. In [50] and [30], the prior knowledge can be interpreted as a specific sensor that is guaranteed to be attacked. In this thesis, we address the prior knowledge from a different perspective; more specifically, if we are given a specific subset of sensors and/or actuators that are not accessible by the attacker, then how can we achieve attack detection and attack correction? What benefit can we gain from the prior attack knowledge in terms of attack detectability and correctability? All those questions are under development. This thesis provides a novel approach to address such issues.

- How to capture the vulnerability of the system in terms of different types of attacks (actuator attack, actuator/sensor attack) using representation-free system parameters is an open research question. In this thesis, we extend the scope of the security index to address different attack scenarios.

In this thesis, our objective is to address the above identified gaps in the literature.

Chapter 3

Linear Cyber-Physical System Security - Sensor Attacks in the Noise-Free Case

3.1 Introduction and Preliminaries

This chapter presents our method for solving the problem of attack detection and attack correction for multi-input multi-output discrete-time linear time-invariant dynamical systems under sensor only attacks. In this chapter, the sensor attack signal is modelled as an additive signal without any restriction (e.g., the power, shape, statistical of the attacks etc.). The only limitation that we impose on the attack signal is the number of sensors being attacked. More specifically, we assume that an upper bound on the number of sensors being attacked. A similar assumptions holds true throughout the remainder of the thesis, apart from Chapter 6, where we consider sensor attacks and bounded noise. In this chapter, we assume the inputs of the system are known, and the attacked sensor outputs are received and measured.

In Section 3.2, we address the problem of attack detection and attack correction without prior sensor knowledge where the attacker potentially has access to all the sensor signals, i.e., any sensors could be attacked and we do not know which ones are attacked.

In this thesis, we choose to use a behavioural approach [46] as a starting point for the discussion. A behavioural approach is an elegant framework to discuss dynamical systems, in particular linear dynamical systems. Based on the input/output behaviour of the system, we recall two types of system representation, the kernel representation and

the Input/Latent/Output (ILO) image representation.

Algorithms for sensor attack detection and correction are presented. The algorithms we propose are guaranteed to achieve attack detection and attack correction so long as certain upper bounds on the number of attacked sensors hold. The upper bounds are stated in terms of the 'sensor security index'. A numerical example that involves a multi-input multi-output dynamical system is presented to illustrate the theory. Some of the main results in Section 3.2 have been published in our papers [62],[61] and [63].

In Section 3.3, we address the problem of attack detection and correction under the assumption that there are sensors that the attacker cannot attack and we know which sensors these are. Such an assumption is represented as a prior sensor knowledge.

In this chapter, we start with representing the system using a kernel representation [46]. Let the attack-free system model be given by its kernel representation:

$$\Sigma : R(\sigma)w = 0 \tag{3.1}$$

where $w = \begin{bmatrix} y \\ u \end{bmatrix}$. The signal $y : \mathbb{Z}_+ \to \mathbb{R}^m$ is the (sensor) output. The signal $u : \mathbb{Z}_+ \to \mathbb{R}^q$ is the known open-loop (actuator) input. The matrix $R(\xi)$ is a full rank real polynomial matrix of size $m \times (m+q)$. In what follows, it matters to consider sub-matrices in $R(\xi)$ as follows:

$$R(\xi) = \begin{bmatrix} R_1(\xi) & R_2(\xi) \end{bmatrix},$$

where the polynomial matrix $R_1(\xi)$ is a full rank $m \times m$ matrix, which implies that the polynomial matrix $R_1(\xi)$ has a non-zero polynomial determinant. Thus Equation (3.1) may be written as

$$R_1(\sigma)y = -R_2(\sigma)u. \tag{3.2}$$

The kernel representation (see for example [46]) we mentioned above is a very general system representation. In later sections, we will see that the kernel representation is compatible with other system representations, for example, state-space representation; image representations. The kernel representation describes a system's dynamics, using

only the input and output signals. It is a kind of minimalistic representation in that sense. Moreover, the attack scenarios being considered in this thesis also fit into such kernel representation since we are considering actuator (input) attacks and sensor (output) attacks.

We define the behaviour of the system as follows:

Definition 3.1. *[46] The behaviour of the system Σ is defined as the signal set given by*

$$\mathcal{B} = \left\{ \begin{bmatrix} y \\ u \end{bmatrix} : \mathbb{Z}_+ \to \mathbb{R}^{m+q} \mid R_1(\sigma)y = -R_2(\sigma)u \right\}, \tag{3.3}$$

where $R_1(\xi)$ and $R_2(\xi)$ are polynomial matrices in the shift operator of dimension $m \times m$ and $m \times q$, respectively. The polynomial matrix $R_1(\xi)$ is full rank.

Thus the behaviour of the system is the set that contains all possible input/output pairs of the system.

The behavioural approach and the kernel representation play vital roles in this thesis. In this setup, we express the dynamics of a system as a subset of time-trajectories. The behavioural approach and kernel representation allows us to present compact and representation-free concepts with regards to the system vulnerability against attacks. It also allows us to propose concise algorithms for attack detection and attack correction.

As we assume that the input signal of the system is known, the homogeneous behaviour is of interest:

Definition 3.2. *[46, Theorem 3.3.19] The homogeneous behaviour of a system Σ is defined as the subset of $\mathcal{B}$ given by*

$$\mathcal{B}_{hom} = \left\{ \begin{bmatrix} y \\ 0 \end{bmatrix} : \mathbb{Z}_+ \to \mathbb{R}^{m+q} \;\middle|\; \begin{bmatrix} y \\ 0 \end{bmatrix} \in \mathcal{B} \right\}. \tag{3.4}$$

Equivalently,

$$\mathcal{B}_{hom} = \left\{ \begin{bmatrix} y \\ 0 \end{bmatrix} : \mathbb{Z}_+ \to \mathbb{R}^{m+q} \;\middle|\; R_1(\xi)y = 0 \right\}. \tag{3.5}$$

Remark: The homogeneous behaviour of the system is a subset of the entire behaviour when the input signal is zero.

We now introduce the following definitions and a lemma that will be used later on in the thesis.

Definition 3.3. *(e.g., Chapter 2, [46]) A square polynomial matrix is called **unimodular** if it has a polynomial inverse.*

Equivalently, a square polynomial matrix is unimodular if its determinant is a non-zero constant.

Definition 3.4. *A polynomial matrix is called **left unimodular** if it has a polynomial left inverse.*

Definition 3.5. *(e.g., Chapter 2.5.5, [46]) Two polynomial matrices $R(\xi)$ and $Q(\xi)$ of the same size are called **left unimodularly equivalent** if there exists a unimodular matrix $U(\xi)$ such that $Q(\xi) = U(\xi)R(\xi)$.*

Lemma 3.1. *[e.g., Theorem 3.9, [38]] Consider two systems Σ and Σ' whose behaviours $\mathcal{B}$ and $\mathcal{B}'$ are given by $\mathcal{B} = \left\{ \begin{bmatrix} y \\ u \end{bmatrix} \middle| R(\sigma) \begin{bmatrix} y \\ u \end{bmatrix} = 0 \right\}$ and $\mathcal{B}' = \left\{ \begin{bmatrix} y' \\ u' \end{bmatrix} \middle| R'(\sigma) \begin{bmatrix} y' \\ u' \end{bmatrix} = 0 \right\}$, respectively. Assume that $R(\xi)$ and $R'(\xi)$ are full rank polynomial matrices of the same size, then $\mathcal{B} = \mathcal{B}'$ if and only if $R(\xi)$ and $R^{'}(\xi)$ are left unimodularly equivalent.*

Thus the above lemma means that left multiplying $R(\xi)$ by a unimodular matrix on system $R(\xi)$ does not change the behaviour of the system.

Now, consider a class of additive sensor attack signals η_s leading to an attacked received signal

$$r = \begin{bmatrix} r_s \\ u \end{bmatrix} = \begin{bmatrix} y \\ u \end{bmatrix} + \begin{bmatrix} \eta_s \\ 0 \end{bmatrix}, \text{ where } \begin{bmatrix} y \\ u \end{bmatrix} \in \mathcal{B}. \tag{3.6}$$

The detectability and correctability of the attack signal η_s regarding system Σ can be defined as follows:

Definition 3.6. *A non-zero attack signal $\begin{bmatrix} \eta_s \\ 0 \end{bmatrix}$ is detectable if $\begin{bmatrix} \eta_s \\ 0 \end{bmatrix} \notin \mathcal{B}_{hom}$.*

From Definition 3.6, it can be seen that an undetectable attack is an attack signal that lies inside the behaviour of the system. The result of such undetectable attack can leads to a different output signal compared with the attack-free output without being detected.

With regard to attack correctability, we first introduce the notion of the attack set $\mathcal{A}$ where we denote $\mathcal{A}$ as the set which formed by all possible attack signals. Now introduce the following definition regarding correctability.

Definition 3.7. *A non-zero attack signal* $\begin{bmatrix} \eta_s \\ 0 \end{bmatrix} \in \mathcal{A}$ *is correctable if for all* $\eta_s' \neq \eta_s$*, the following is satisfied*

$$\begin{bmatrix} \eta_s' \\ 0 \end{bmatrix} \in \mathcal{A} \Rightarrow \begin{bmatrix} \eta_s \\ 0 \end{bmatrix} - \begin{bmatrix} \eta_s' \\ 0 \end{bmatrix} \notin \mathcal{B}_{hom}. \tag{3.7}$$

The essence of the correctability definition is that if an attack signal is correctable, then there exists a unique way in which to correct the attack.

As we mentioned in Section 2.2, attacks are more dangerous than faults. From the definition of detectability and correctability, it is not hard to see that attacks are more dangerous compared with faults. For example, a smart attack that mimic the system's behaviour at a certain level can be undetectable or uncorrectable, while fault signals do not inherent such property.

The objectives in the following sections are first determine the conditions for a successful detection and a unique correction. We then propose attack detection as well as attack correction methods.

3.2 Sensor Attack Without Prior Sensor Knowledge

In this section, we present our sensor attack detection/correction methods without prior sensor knowledge. Recall the previous chapter, a prior sensor knowledge can be interpreted as: the user knows beforehand that a specific sensor (or a specific subset of sensors) is guaranteed to be attacked (or attack-free). In this section, we do not make such an assumption, meaning any of the sensor signals is potentially under attack.

3.2.1 Security Index and Input/Latent/Output Image Representation

In this subsection, we first recall the concept of the sensor security index. Such concept was first introduced for systems with only outputs in [15]. In this subsection, we ex-

tend the this concept to systems with outputs and known inputs. We published this in [61]. The sensor security index characterises the vulnerability of a system against sensor only attacks. We then present an equivalent system representation, namely the Input/Latent/Output (ILO) image representation. The ILO image representation is a useful tool when discussing the attack correction method.

Sensor security index

The vulnerability of a system (3.3) against sensor attack with no prior sensor knowledge is addressed using the notion of sensor security index $\delta_s(\Sigma)$ of [14]. The subscript 's' stands for sensor attack. In [14, 15], the sensor security index is proposed for systems without input; in [61] we extended this concept for systems with known inputs.

Definition 3.8. *The sensor security index for system Σ is defined as follows:*

$$\delta_s(\Sigma) := \min_{0 \neq \begin{bmatrix} y \\ 0 \end{bmatrix} \in \mathcal{B}_{hom}} \|y\|. \tag{3.8}$$

Further define $\delta_s(\Sigma) = m + 1$ if $\mathcal{B}_{hom} = \{0\}$.

Definition 3.9. *We call a system Σ with m outputs **trivially secure** subject to sensor attack if $\delta_s(\Sigma) = m + 1$.*

In [61, 62], the terminology 'maximally secure' was used when $\delta_s(\Sigma) = m$; here the new concept of 'trivially secure' was used when $\delta_s(\Sigma) = m + 1$. The difference is that when a system is trivially secure, then the output is uniquely determined by the known input and the CPS security issue (attack detection and correction) can be solved trivially, as will be shown later on in this thesis.

It is difficult to compute $\delta_s(\Sigma)$ from the definition, because we need to search the entire $\mathcal{B}_{hom}$. The following theorem provides a more appropriate method to calculate $\delta_s(\Sigma)$ based on the kernel representation of system Σ as in (3.2).

Theorem 3.1. *Consider a system* Σ *given by* (3.2)*, where* $R_1(\xi)$ *has full rank, then*

$$\delta_s(\Sigma) = k + 1, \tag{3.9}$$

where k is the largest integer in $\{1, ..., m\}$ *such that for any subset* $\mathcal{J} \subseteq \{1, \ldots, m\}$ *of cardinality k, the* $m \times k$ *matrix* $R_1^{(\bullet,\mathcal{J})}(\xi)$ *is left unimodular.*

Proof. First consider the trivially secure case where the only solution for equation $R_1(\sigma)y = 0$ is $y = 0$. This implies that the polynomial matrix $R_1(\xi)$ is unimodular. We then have $k = m$ and $\delta_s(\Sigma) = m + 1$, so that equation (3.9) holds.

For systems that are not trivially secure, i.e., $k \leq m - 1$, clearly there exists a subset $\mathcal{J} \subseteq \{1, \ldots, m\}$ of cardinality $k + 1$ such that $R^{(\bullet,\mathcal{J})}(\xi)$ is not left unimodular. Thus there exists a non-zero signal $y^\star$ that satisfies $R^{(\bullet,\mathcal{J})}(\xi)y^\star = 0$. Now let $y : \mathbb{Z}_+ \to \mathbb{R}^m$ be the signal satisfying $y^{(\mathcal{J})} = y^\star$ and $y^{(\bar{\mathcal{J}})} = 0$. Then $\begin{bmatrix} y \\ 0 \end{bmatrix} \in \mathcal{B}_{hom}$ and $\|y\| = \|y^\star\| \leq k + 1$. This implies that $\delta_s(\Sigma) \leq k + 1$.

To prove that also $\delta_s(\Sigma) \geq k + 1$, let $\begin{bmatrix} y \\ 0 \end{bmatrix}$ be a signal in $\mathcal{B}_{hom}$ of weight $\delta_s(\Sigma)$. Define $\bar{\mathcal{J}} \subset \{1, 2, \ldots, m\}$ as a set of cardinality $\delta_s(\Sigma)$ such that $y^{(\mathcal{J})} = 0$. Then $R^{(\bullet,\bar{\mathcal{J}})}(\xi)y^{(\bar{\mathcal{J}})} = 0$ and because $y^{(\bar{\mathcal{J}})} \neq 0$ it follows that $R^{(\bullet,\bar{\mathcal{J}})}(\xi)$ is not left unimodular. This implies that $\delta_s(\Sigma) > k$. This completes the proof. □

Remark: To check the left unimodularity of a polynomial matrix $R_1^{(\bullet,\mathcal{J})}(\xi)$, one can use e.g. Theorem B.1.1 in [46]. First, transform $R_1^{(\bullet,\mathcal{J})}(\xi)$ into an upper triangular polynomial matrix (see e.g., Theorem 2.5.14, [46]), and then check the determinant of the top $\|\mathcal{J}\|_c \times \|\mathcal{J}\|_c$ submatrix. If the determinant is a non-zero constant, then the original polynomial matrix $R_1^{(\bullet,\mathcal{J})}(\xi)$ is left unimodular.

Remark: Based on Theorem 3.1, in the trivially secure case, it follows that $R_1(\xi)$ is unimodular, then without loss of generality, the dynamics of a trivially secure system can be expressed as follows:

$$y = R_2(\sigma)u. \tag{3.10}$$

The following theorems provide sufficient conditions for attack detectability and correctability using the notion of $\delta_s(\Sigma)$.

Theorem 3.2. *If a non-zero sensor attack signal η_s satisfies $\|\eta_s\| < \delta_s(\Sigma)$, then $\begin{bmatrix} \eta_s \\ 0 \end{bmatrix}$ is a detectable attack signal.*

Proof. For any non-zero attack signal $\begin{bmatrix} \eta_s \\ 0 \end{bmatrix}$ satisfying $\|\eta_s\| < \delta_s(\Sigma)$, we must have $\begin{bmatrix} \eta_s \\ 0 \end{bmatrix} \notin \mathcal{B}_{hom}$ because of Definition 3.8. According to Definition 3.6, η_s is then detectable. □

Theorem 3.3. *Let the attack set $\mathcal{A}$ be defined as:*

$$\mathcal{A} = \left\{ \begin{bmatrix} \eta_s \\ 0 \end{bmatrix} : \mathbb{Z}_+ \to \mathbb{R}^{m+q} \;\middle|\; 0 \neq \|\eta_s\| < \delta_s(\Sigma)/2 \text{ and } \eta_s : \mathbb{Z}_+ \to \mathbb{R}^m \right\}.$$

Then all attack signals $\begin{bmatrix} \eta_s \\ 0 \end{bmatrix} \in \mathcal{A}$ are correctable.

Proof. Consider an attack signal $\begin{bmatrix} \eta_s \\ 0 \end{bmatrix}$ from $\mathcal{A}$. If there exists another non-zero signal $\begin{bmatrix} \eta_s' \\ 0 \end{bmatrix}$ from $\mathcal{A}$ with $\|\eta_s'\| < \delta_s(\Sigma)/2$ and $\eta_s' \neq \eta_s$, then $\|\eta_s - \eta_s'\| \leq \|\eta_s\| + \|\eta_s'\| < \delta_s(\Sigma)$ which implies $\begin{bmatrix} \eta_s - \eta_s' \\ 0 \end{bmatrix} \notin \mathcal{B}_{hom}$. According to Definition 3.7, $\eta_s \in \mathcal{A}$ is then correctable and this completes the proof. □

The threshold $\delta_s(\Sigma)/2$ provides a sufficient condition for an attack signal being correctable, which implies that if the number of sensors being attacked is smaller than $\delta_s(\Sigma)/2$, then a unique correction can be found in the attack set $\mathcal{A}$. However, when the attack set $\mathcal{A}$ is not bounded by $\delta_s/2$, then there may exists multiple signals η_s' such that $\begin{bmatrix} \eta_s - \eta_s' \\ 0 \end{bmatrix} \in \mathcal{B}_{hom}$, this implies that unique correction is not possible.

Unlike [13], [56], in which the vulnerability of a system against attacks is characterised based on the system representation, the sensor security index $\delta_s(\Sigma)$ is defined

as a representation-free concept. This means it is a system property independent of its representation. In the results of [13] and [56] we notice that the representation of the system is vital in terms of detecting and/or correcting an attack signal. In our work, this is not found to be the case. Detection and correction are defined in a representation-free form, such representation-free concept provides a general vulnerability measurement for systems under adversarial attacks. More specifically, such general representation-free concept can be described under various system representations. For example, in [14], the authors discussed how to calculate such sensor security index not only in kernel representation, but also state-space representation.

In Section B of [30], and [51], the notion of security index α_i for sensor i was presented, which expresses the vulnerability of sensor i against adversarial attacks for system Σ. Recall equation (2.33), the concept of α_i can be summarised as follows: if an attacker wants to attack the i-th received sensor signal, then how many other sensors should also be attacked in order to remain undetectable? The difference between $\delta_s(\Sigma)$ and α_i is that $\delta_s(\Sigma)$ focuses on describing the vulnerability of the overall system. While the focus for α_i is on individual sensor. If a noise-free system is considered in [30] and [51], and further assume α_i is finite for all i (in [30,51], the value of α_i is define to be infinite if there exists no such undetectable attack signal for sensor i, in our scope, this is equivalent to the trivially secure case), then the following Definition and Lemma explains the relationship between α_i and $\delta_s(\Sigma)$.

Definition 3.10. *Consider a system Σ in a behavioural approach as in equation* (3.2)*, α_i can be defined as*

$$\alpha_i = \min_{\begin{bmatrix} y \\ 0 \end{bmatrix} \in \mathcal{B}, y^{(i)} \neq 0} \|y^{(\bar{i})}\|,$$

Lemma 3.2. *For a system Σ that is not trivially secure, the following equation must hold:*

$$\delta_s(\Sigma) - 1 = \min_{i \in \{1,\dots,m\}} \alpha_i.$$

Proof. Given

$$\alpha_i = \min_{\begin{bmatrix} y \\ 0 \end{bmatrix} \in \mathcal{B},\, y^{(i)} \neq 0} \|y^{(\bar{i})}\|,$$

then

$$\alpha_i + 1 = \min_{\begin{bmatrix} y \\ 0 \end{bmatrix} \in \mathcal{B},\, y^{(i)} \neq 0} \|y\|.$$

Define

$$\mathcal{B}_i := \left\{ \begin{bmatrix} y \\ 0 \end{bmatrix} \;\middle|\; \begin{bmatrix} y \\ 0 \end{bmatrix} \in \mathcal{B},\, y^{(i)} \neq 0 \right\},$$

and we have

$$\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2 ... \cup \mathcal{B}_m.$$

Given α_i is finite for all i, we then have

$$\min_{i \in \{1,...,m\}} \alpha_i + 1 = \min_{0 \neq \begin{bmatrix} y \\ 0 \end{bmatrix} \in \mathcal{B}} \|y\| = \delta_s(\Sigma),$$

this completes the proof. □

ILO image representation

Let us now consider a different system representation. We first recall Lemma 3.1, when we describe a system's behaviour $\mathcal{B}$, the polynomial matrix $R(\xi)$ is not unique. Among all the different matrices $R(\xi)$ describing a same behaviour, we single out a canonical form where $R_1(\xi)$ is represented in the Kronecker-Hermite canonical form [28],[18],[33]. Such particular form plays an important role in the sequel. It also provides connections for a class of system representations, namely, Input/Latent/Output (ILO) image representations. The following results are also in our papers [61, 62].

Theorem 3.4. *Consider a system* Σ *given by* (3.3) *with sensor security index* $\delta_s(\Sigma)$*, whose ho-*

mogeneous behavior $\mathcal{B}_{hom}$ is given by (3.4) *and* (3.5)*, where $R_1(\xi)$ has full rank, then there exists a unimodular matrix $U(\xi)$ such that*

$$U(\xi)R_1(\xi) = \left[\begin{array}{c|c} -\mathbb{I}_{\delta_s(\Sigma)-1} & M_1(\xi) \\ \hline 0 \quad \dots \quad 0 & D(\xi) \end{array}\right]. \tag{3.11}$$

where $D(\xi)$ is an upper triangular matrix and the degree of the diagonal entities of $D(\xi)$ denoted as $deg\{d_{ii}(\xi)\}$ for $i \in \{1, ..., m - \delta_s(\Sigma) + 1\}$, is strictly the highest within the corresponding column of (3.11)*. In particular, the system is represented in the following Kronecker-Hermite canonical form:*

$$\left[\begin{array}{c|c} \mathbb{I}_{\delta_s(\Sigma)-1} & -M_1(\sigma) \\ \hline 0 \quad \dots \quad 0 & -D(\sigma) \end{array}\right] y = R_2(\sigma)u. \tag{3.12}$$

.

Proof. It follows from Theorem B.1.1 in [46] that there exists a unimodular matrix $U_0(\xi)$ such that $U_0(\xi)R_1(\xi)$ is an upper triangular polynomial matrix, say $R_0(\xi)$, written as

$$R_0(\xi) = \begin{bmatrix} a_1(\xi) & b_{12}(\xi) & \dots & b_{1(m-1)}(\xi) & b_{1m}(\xi) \\ 0 & a_2(\xi) & \dots & b_{2(m-1)}(\xi) & b_{2m}(\xi) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & a_{(m-1)}(\xi) & b_{(m-1)m}(\xi) \\ 0 & 0 & \dots & 0 & a_m(\xi) \end{bmatrix}.$$

By Theorem 3.1, since all $m \times (\delta_s(\Sigma) - 1)$ submatrices of $R_1(\xi)$ are left unimodular, it follows that in particular, the matrix is formed by the first $\delta_s(\Sigma) - 1$ columns is left unimodular. This implies that all its diagonal elements for those columns are now non-zero constants. Without restrictions, $R_0(\xi)$ can then be written as

$$\begin{bmatrix} 1 & b_{12}(\xi) & \dots & b_{1(\delta_s(\Sigma)-1)}(\xi) & \dots & b_{1m}(\xi) \\ 0 & 1 & \dots & b_{2(\delta_s(\Sigma)-1)}(\xi) & \dots & b_{2m}(\xi) \\ \vdots & \vdots & \ddots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 & \dots & b_{(\delta_s(\Sigma)-1)m}(\xi) \\ \vdots & \vdots & \dots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & \dots & a_m(\xi) \end{bmatrix}. \tag{3.13}$$

We can now apply classical theory see e.g., [28] Theorem 2.40, [18] Theorem 7.5 or [33] to transform the matrix in (3.13) into the form (3.11) using left multiplication by a unimodular matrix. It is a classical result that there exists a unimodular matrix $U_1(\xi)$ such that $U_1(\xi)R_0(\xi)$ is in Kronecker-Hermite canonical form as in (3.11) with $D(\xi)$ is an upper triangular matrix and that the degree of the diagonal entities of $D(\xi)$ is strictly the highest within the corresponding column, see e.g., [28, Theorem 2.40], [18, Theorem 7.5] or [33].

When a system is trivially secure, i.e., $\delta_s(\Sigma) = m + 1$, the theorem also holds. In this case, the matrix $\begin{bmatrix} M_1(\xi) \\ D(\xi) \end{bmatrix}$ does not exist. Thus, the Kronecker-Hermite canonical form of R_1 is the identity matrix. This coincides with equation (3.10). More specifically, we denote the ILO image representation when the system is trivially secure as follows:

$$\begin{bmatrix} y \\ 0 \end{bmatrix} = \begin{bmatrix} P(\sigma) \\ Q(\sigma) \end{bmatrix} u. \tag{3.14}$$

□

For systems that are not trivially secure, the equation in representation (3.12) can be rewritten in the following form:

$$\begin{bmatrix} y \\ 0 \end{bmatrix} = \begin{bmatrix} M(\sigma) & P(\sigma) \\ D(\sigma) & Q(\sigma) \end{bmatrix} \begin{bmatrix} l \\ u \end{bmatrix}, \tag{3.15}$$

where the signal l is a latent signal that shows similar properties as a 'state signal'. We will explain the intuition behind such interpretation later in this section. In this case, the signal l simply coincides with the last $m - \delta_s(\Sigma) + 1$ components of the output signal y because matrix $M(\xi) = \begin{bmatrix} M_1(\xi) \\ \mathbb{I}_{m-\delta_s(\Sigma)+1} \end{bmatrix}$ is a polynomial matrix of size $m \times (m - \delta_s(\Sigma) + 1)$. The matrix $P(\xi) = \begin{bmatrix} R_2^*(\xi) \\ 0 \end{bmatrix}$ where $R_2^*(\xi)$ is the first $\delta_s(\Sigma) - 1$ rows of $R_2(\xi)$, the size of $R_2^*(\xi)$ is $m \times (m - \delta_s(\Sigma) + 1)$. The polynomial matrix $Q(\xi)$ is formed by taking the last $m - \delta_s(\Sigma) + 1$ rows of $R_2(\xi)$.

From the above theorem, we can gain more insight regarding the full-rankness as-

sumption for $R_1(\xi)$. In fact, such assumption is related to the concept of 'observability' as in Definition 2.1. Recall a system is said to be observable if for any given u, the state signal x can be determined in finite time using only y. Under the scope, we say the system is observable if for any given u, the latent signal l can be determined in finite time using only y. In the proposed ILO image representation, the full rankness of $R_1(\xi)$ guarantees the latent signal l is observable from the output y for any given u.

We now propose the following theorem regarding the minimum number of latent signals for the ILO image representation (3.15).

Theorem 3.5. *Consider a system Σ in its ILO image representation as equation* (3.15) *with m outputs and the sensor security index equals $\delta_s(\Sigma)$, then the minimum number of latent signals required to describe the behaviour $\mathcal{B}$ of Σ is $m - \delta_s(\Sigma) + 1$.*

Proof. If there exists another latent signal l' with the size of l' is a number $< m - \delta_s(\Sigma) + 1$ that also describes the same input/output behaviour $\mathcal{B}$, then such l' also describes the same homogenous behaviour $\mathcal{B}_{hom}$. Set the input $u = 0$, we then have the following relationship:

$$\begin{bmatrix} M(\sigma) \\ D(\sigma) \end{bmatrix} l = \begin{bmatrix} M'(\sigma) \\ D'(\sigma) \end{bmatrix} l'.$$

If the above equation holds and the size of l' smaller than $< m - \delta_s(\Sigma) + 1$, then the column rank of $\begin{bmatrix} M(\sigma) \\ D(\sigma) \end{bmatrix}$ must be smaller than $< m - \delta_s(\Sigma) + 1$. Since $M(\xi) = \begin{bmatrix} M_1(\xi) \\ \mathbb{I}_{m-\delta_s(\Sigma)+1} \end{bmatrix}$ and $\begin{bmatrix} M(\sigma) \\ D(\sigma) \end{bmatrix}$ has full column rank, which is a contradiction. Thus we can conclude that the minimum size of l is $m - \delta_s(\Sigma) + 1$. □

Based on Theorem 3.5, it can be seen that the minimal number of latent signals in need for representing a system's dynamics is related to the concept of the sensor security index. When we are given the number of outputs is m, it is observed that the larger the security index, the fewer the latent signals. From another point of view, a larger security index means the system has more sensor redundancy since for a fixed number of outputs, and the system has less latent variables. On the other hand, a smaller security index implies

that the system has less sensor redundancy since there are many latent signals and fewer sensor measurements.

However, in general, the number of components for l can be larger than $m - \delta_s(\Sigma) + 1$. In the rest of this thesis, unless otherwise stated, l is not restricted to be minimal, and we denote the size of l by n, i.e., $l : \mathbb{Z}_+ \to \mathbb{R}^n$.

Based on equation (3.15), a observable system under sensor attack can be expressed as follows:

$$\begin{bmatrix} r_s - \eta_s \\ 0 \end{bmatrix} = \begin{bmatrix} M(\sigma) & P(\sigma) \\ D(\sigma) & Q(\sigma) \end{bmatrix} \begin{bmatrix} l \\ u \end{bmatrix}. \tag{3.16}$$

We denote the system representation (3.15) as an Input/Latent/Output (ILO) image representation. In Chapter 4 of [46], the authors present the concept of Input/State/Output model, or ISO model. Our ILO image representation shows similarities with this well-known ISO model, but the fundamental differences also exist in many aspects. In the following paragraphs, we will explain in detail regarding the similarities and the differences between these two system representations.

The starting point for the ISO model in [46] is a state-space representation as in equation (2.2), namely:

$$\begin{aligned} \sigma x &= Ax + Bu \\ y &= Cx + Du. \end{aligned} \tag{3.17}$$

In polynomial form, equation (3.17) can be written as the following ISO model:

$$\tilde{R}(\sigma) \begin{bmatrix} u \\ y \end{bmatrix} = \tilde{M}(\sigma) x.$$

Such ISO model is very general, in a sense that there exists a direct relationship between the state space representation and the ISO model:

$$\tilde{R}(\xi) = \begin{bmatrix} B & 0 \\ -D & \mathbb{I} \end{bmatrix}, \quad M(\xi) = \begin{bmatrix} \xi\mathbb{I} - A \\ C \end{bmatrix}.$$

Our proposed ILO image representation (equation (3.15)) exhibits a similar property. In fact, a state-space representation can be treated as a special case of (3.15). In our case, the state-space representation (3.17) can also be treated as a special case of (3.15), namely, by letting $M(\xi) = C, D(\xi) = \xi\mathbb{I} - A, P(\xi) = D$ and $Q(\xi) = -B$.

In both ISO model in [46], the variable x is considered as a latent variable. In terms of the input/output behaviour of the system, the state variable in ISO model or the latent signal in ILO image representation plays a similar role: influenced by the input u and drives the output y.

However, the focus of the ILO image representation and the ISO model is different. In this chapter, we are dealing with systems under sensor attacks, while the input signals are attack free, and the latent signals are not corrupted. Thus it is vital to present an image representation that isolates the output signals on one side of the equation. Later in this chapter, we will see that such an ILO image representation allow us to isolate a subset of corrupted outputs. On the other hand, the focus for the ISO model in [46] is to describe not only the entire behaviour of the system but also the input/state and the output/state behaviour.

Using the ILO image representation, the behaviour $\mathcal{B}$ of the system Σ can be expressed as:

$$\mathcal{B} = \left\{ \begin{bmatrix} y \\ u \end{bmatrix} : \mathbb{Z}_+ \to \mathbb{R}^{m+q} \,\middle|\, \begin{bmatrix} y \\ 0 \end{bmatrix} = \begin{bmatrix} M(\sigma) & P(\sigma) \\ D(\sigma) & Q(\sigma) \end{bmatrix} \begin{bmatrix} l \\ u \end{bmatrix} \text{ for some } l : \mathbb{Z}_+ \to \mathbb{R}^n \right\}. \quad (3.18)$$

From Theorem 3.4, we can conclude that for any system expressed in a kernel representation with $R_1(\xi)$ full rank, there exists an ILO image representation that characterises the same input-output behaviour. It is natural to ask that if an ILO image representation is used to characterise the behaviour of a observable system, can we find a kernel representation to achieve the same purpose? If the statement is true, then we can conclude that the kernel representation and the ILO image representation are equivalent. In fact, similar statements have proven to be correct (see Chapter 6.4 of [46] and also [71], [72]). In

the remainder of this sub-section, we review those approaches, and provide explanations based on a check matrix $H(\xi)$. In order to achieve this, we first present the following theorem:

Theorem 3.6. *Consider a observable system in its ILO image representation as in* (3.15)*, then there exists an* $m \times (n+m)$ *polynomial matrix* $H(\xi)$ *such that*

$$ker\{H(\xi)\} = image \left\{ \begin{bmatrix} M(\xi) \\ D(\xi) \end{bmatrix} \right\}$$

where ker$(\bullet)$ *denotes the kernel of the matrix* $(\bullet)$ *while image*$(\bullet)$ *represents the image of the matrix* $(\bullet)$*.*

Proof. Given the system is observable, there exists a polynomial left inverse of $\begin{bmatrix} M(\xi) \\ D(\xi) \end{bmatrix}$ (see Chapter 5.3 in [46]). This is equivalent to say that there exists a unimodular matrix $U(\xi)$ of size $(m+n) \times (m+n)$ such that

$$U(\xi) \begin{bmatrix} M(\xi) \\ D(\xi) \end{bmatrix} = \begin{bmatrix} \mathbb{I}_n \\ 0 \end{bmatrix}. \tag{3.19}$$

Define the polynomial matrix $H(\xi)$ as follows:

$$H(\xi) = \begin{bmatrix} 0_{m \times n} & \mathbb{I}_m \end{bmatrix} U(\xi), \tag{3.20}$$

we have

$$H(\xi) \begin{bmatrix} M(\xi) \\ D(\xi) \end{bmatrix} = 0.$$

It can also be seen that polynomial matrix $H(\xi)$ has a full row rank. Since $\begin{bmatrix} M(\xi) \\ D(\xi) \end{bmatrix}$ has full column rank and moreover:

$$\text{The number of columns of } H(\xi) = \text{row rank of } H(\xi) + \text{column rank of } \begin{bmatrix} M(\xi) \\ D(\xi) \end{bmatrix},$$

then we must have the kernel of $H(\xi)$ equals the image of $\begin{bmatrix} M(\xi) \\ D(\xi) \end{bmatrix}$, i.e.,

$$\ker\{H(\xi)\} = \text{image}\left\{ \begin{bmatrix} M(\xi) \\ D(\xi) \end{bmatrix} \right\}.$$

This then completes the proof of the theorem. □

Remark: In this thesis, we call the polynomial matrix $H(\xi)$ the **check matrix** since it exhibits similarities with the parity check matrix in coding theory [47].

Based on Theorem 3.6, we now present Theorem 3.7 around an equivalent system representation.

Theorem 3.7. *For any observable system expressed in the ILO image representation as in* (3.15) *with behaviour $\mathcal{B}$ as in equation* (3.18)*, there exists a kernel representation as in* (3.1) *that express the same behaviour $\mathcal{B}$ as in equation* (3.3).

Proof. Recall that the system is given by Equation (3.15),

$$\begin{bmatrix} M(\sigma) \\ D(\sigma) \end{bmatrix} l = \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} P(\sigma) \\ Q(\sigma) \end{bmatrix} u = \begin{bmatrix} \mathbb{I} & -P(\sigma) \\ 0 & -Q(\sigma) \end{bmatrix} \begin{bmatrix} y \\ u \end{bmatrix}.$$

Recall Theorem 3.6 and denote $R(\xi) = H(\xi) \begin{bmatrix} \mathbb{I} & -P(\xi) \\ 0 & -Q(\xi) \end{bmatrix}$, we then have $R(\sigma) \begin{bmatrix} y \\ u \end{bmatrix} = 0$ as in (3.1). This completes the proof. □

Based on Theorem 3.4 and Theorem 3.7, we can conclude that the kernel representation (3.1), (3.2) and the ILO image representation (3.15) are equivalent in that they describe the same behaviour. For this reasoning, both kernel/ILO image representations may be used for discussing different problems in the following sections and chapters. Notably, the ILO image representation (3.15) is instrumental in our derivation of the attack correction method in this section, and is also vital to the following chapters in this thesis.

As the next step towards attack detection and correction, we now express the system's sensor security index in terms of the polynomial matrices $M(\xi)$ and $D(\xi)$.

Theorem 3.8. *Consider a system Σ that is not trivially secure and given by* (3.15)*, then*

$$\delta_s(\Sigma) = m + 1 - k, \tag{3.21}$$

where k is the smallest integer in $\{0, ..., m\}$ such that for any subset $\mathcal{J} \subseteq \{1, \dots, m\}$ of cardinality k, the polynomial matrix $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) \\ D(\xi) \end{bmatrix}$ is left unimodular.

Proof. Consider the homogeneous behaviour of the system by setting input signal $u = 0$:

$$\begin{bmatrix} y \\ 0 \end{bmatrix} = \begin{bmatrix} M(\sigma) \\ D(\sigma) \end{bmatrix} l, \text{ where } \begin{bmatrix} y \\ 0 \end{bmatrix} \in \mathcal{B}_{hom}. \tag{3.22}$$

For systems that are not trivially secure, clearly, there exists a subset $\mathcal{J} \subseteq \{1, \dots, m\}$ of cardinality $k-1$ such that $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) \\ D(\xi) \end{bmatrix}$ is not left unimodular. Thus, there exists a non-zero signal $l^\star$ that satisfies $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\sigma) \\ D(\sigma) \end{bmatrix} l^\star = 0$. Now, let us consider the signal y defined as $y := M(\sigma)l^\star$. Clearly $\|y\| \leq m - (k-1) = m - k + 1$. This implies that

$$\delta_s(\Sigma) \leq m - k + 1. \tag{3.23}$$

To prove that also $\delta_s(\Sigma) \geq m - k + 1$, let $y^\star$ be a signal in $\mathcal{B}_{hom}$ of weight $\delta_s(\Sigma)$. Thus there exists a non-zero signal $l^\star$ such that $\begin{bmatrix} \mathbb{I}_m \\ 0 \end{bmatrix} y^\star = \begin{bmatrix} M(\sigma) \\ D(\sigma) \end{bmatrix} l^\star$. Define $\bar{\mathcal{J}} \subset \{1, 2, \dots, m\}$ as the set of cardinality $\delta_s(\Sigma)$ for which $y^{\star(\mathcal{J})} = 0$. Then $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\sigma) \\ D(\sigma) \end{bmatrix} l^\star = 0$ and because $l^\star \neq 0$, it follows that $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) \\ D(\xi) \end{bmatrix}$ is not left unimodular. In turn, this implies that $k \geq m - \delta_s(\Sigma) + 1$ and because of (3.23), the proof is now complete. □

Remark: When a system is trivially secure, the number of columns of $\begin{bmatrix} M(\xi) \\ D(\xi) \end{bmatrix}$ equals zero and in this case Theorem 3.8 is not applicable. In this case, the attack-free system representation can be expressed as $y = P(\sigma)u$. In this case, the sensor security index can be read-off from the system representation.

Remark: Now let $\mathcal{J}$ be a subset of $\{1, ..., m\}$ of cardinality $p = m + 1 - \delta_s(\Sigma)$. Since the matrix $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) \\ D(\xi) \end{bmatrix}$ is left unimodular, there then exists a polynomial matrix $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) \\ D(\xi) \end{bmatrix}^{-1}$ such that

$$\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) \\ D(\xi) \end{bmatrix}^{-1} \begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) \\ D(\xi) \end{bmatrix} = \mathbb{I}_p. \tag{3.24}$$

The polynomial matrix $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) \\ D(\xi) \end{bmatrix}^{-1}$ for each $\mathcal{J}$ can be interpreted as observers that isolate some of the sensor measurements for reconstructing the latent signal. This will be explained in greater detail later on in this chapter.

Based on the previous discussion regarding the sensor security index in an ILO image representation, it can be seen that the concept of the security index is very general. In fact, several other concepts in the existing literature regarding sensor attack security can be reformulated in terms of our notion of the sensor security index $\delta_s(\Sigma)$. More specifically, the work of [13] proposes the concept of M-attack observability of a state-space representation (A, C) in terms of the system matrix A and output matrix C. With regard to the sensor security index, any observable state-space representation (A, C) is $\lfloor \delta_s(\Sigma)/2 \rfloor$-attack observable. In the terminology of [56], we can phrase this as $(\delta_s(\Sigma) - 1)$-sparse attack observability.

3.2.2 Attack Detection

In this section, we propose two attack detection algorithms for dynamical systems with non-zero but known inputs under adversarial sensor attacks. First for kernel representations, then for ILO image representations. In this subsection and also in later chapters of this thesis, we will see that the proposed attack detection algorithms are not only limited

to the sensor only attack case. In fact, those two algorithms are universal attack detection algorithms in the sense that both two algorithms are suitable for any attack scenarios being considered in this thesis. For the sake of consistency in this subsection, we illustrate the attack detection algorithms using sensor only attacks as an example.

Recall the previous section; we define the received signal for the attacked system as:

$$r = \begin{bmatrix} r_s \\ u \end{bmatrix} = \begin{bmatrix} y + \eta_s \\ u \end{bmatrix}. \tag{3.25}$$

In linear coding theory, syndrome computation is an effective method for error detection, see e.g. [47, Chapter 7]. Similarly, we will work with a signal that can be interpreted as a 'residual signal'. We define a residual signal as

$$s = R_1(\sigma)r_s + R_2(\sigma)u. \tag{3.26}$$

Here, the signal r_s is the received output signal that may have been attacked; u is the system's input signal. Both r_s and u are assumed to be known.

The notion of the residual signal is often used in the fault detection literature, for example, [16, 17, 24, 34]. We note that our residual signal fits into the general concept of the residual signal as in (2.4). In both these cases, residual signals are linear combinations of the past of the inputs and the measurements.

Unlike [16, 17, 24, 34], the systems being considered in this chapter are noise-free systems; thus, all measurements are precise. We restrict ourselves to a noise-free environment to develop a conceptual approach to attack detection and correction.

The objective of attack detection is to determine whether or not an attack has occurred. We are now ready to present our attack detection Algorithm 3 and 4.

Attack detection in kernel representation

For systems in kernel representation, we propose attack detection Algorithm 3.

Theorem 3.9. *Consider a system given by* (3.1). *Let* $r = \begin{bmatrix} y + \eta_s \\ u \end{bmatrix}$ *be the received signal with*

Algorithm 3 Attack detection for system Σ in kernel representation (3.2)

1: **procedure** $(R_1(\xi), R_2(\xi), r_s, u, \eta_s)$
   ▷ Given $R_1(\xi), R_2(\xi), r_s$ and u, detect whether η_s is the zero signal.
2: Calculate $s = R_1(\sigma)r_s + R_2(\sigma)u$.
3: **if** $s = 0$ **then** decide no attack, i.e., $\eta_s = 0$.
4: **else** decide attack occurred, i.e., $\eta_s \neq 0$.
5: **end if**
6: **end procedure**

$\begin{bmatrix} y \\ u \end{bmatrix} \in \mathcal{B}$, *then the residual signal* $s = 0$ *if and only if the attack signal* $\begin{bmatrix} \eta_s \\ 0 \end{bmatrix} \in \mathcal{B}_{hom}$. *Thus Algorithm 3 gives the correct result if and only if* η_s *is detectable or* $\eta_s = 0$.

Proof.

$$\begin{aligned} s &= R_1(\sigma)(y + \eta_s) + R_2(\sigma)u \\ &= R_1(\sigma)\eta_s. \end{aligned}$$

The theorem now follows trivially. □

This detection algorithm is similar to a parity relations approach as in, for example, [10,29], the parity relations in can be expressed using the following polynomial matrices:

$$s = H(\sigma)[r_s - N(\sigma)u] = Z(\sigma)\eta_s. \tag{3.27}$$

Where $N(\xi)$ can be interpreted as an input-to-output matrix and $Z(\xi)$ is the equivalent attack (fault) to residual parity matrix. In our case, the residual signal $s = R_1(\sigma)\eta_s$, which has the same structure as in (3.27). The difference is that in our case, the parity relations already lies in the system representation (Using $R_1(\xi)$), while in [29], the parity matrix is computed by solving a group of autoregressive–moving-average (ARMA) equations. As a conclusion, we can see that less effort is needed for finding the parity check matrices in our case.

Comparing this detection algorithm with recent attack detection methods such as Algorithm 1 in Chapter 2 from [42] where a residual signal is defined as $s(t) = r(t) - \tilde{y}(t)$ and $\tilde{y}(t)$ shows an estimated output signal produced by a Kalman filter-based observer.

Our detection algorithm can also be recognised as a residual-based detection algorithm. There are two main differences between our algorithm and the existing methods:

1, in this chapter, we restrict ourselves to a noise-free environment, which enables us to understand the essence of the attack security. Thus, the residual signal is certain. For this reason, the residual signal is generated based from the system dynamics only. No further process with respect to the input/output signal is required (such as generating an estimated state/output signal, etc.).

2, using the notion of signal trajectories, the focus for this detection algorithm is on the signal spaces rather than signal values at particular time instants. By doing so, a concise detection algorithm as in Algorithm 3 can be presented.

From Algorithm 3, we see that the amount of delay equals the highest polynomial degree in $R(\xi)$ in discrete time, which coincides with our design aim. Moreover, we can generate all residual signals using the following equation:

$$s = U(\sigma)R_1(\sigma)r_s + U(\sigma)R_2(\sigma)u,$$

where $U(\xi)$ is any unimodular matrix. Put differently, there exists design freedoms for such a residual signal. This statement shows similarity as was evidenced in [16], where the authors also noticed the parity relations are not unique.

How to reduce the amount of delay for our previously proposed detection algorithm is beyond the scope of this thesis, and this will be one of our future research directions.

Attack detection in ILO image representation

In this subsection, we propose an attack detection algorithm for systems in the ILO image representation.

First let us recall that the attacked system (3.16) and the check matrix $H(\xi)$ as in Theorem 3.6, and define the signal s as follows:

$$s = H(\sigma)\left(\begin{bmatrix} r_s \\ 0 \end{bmatrix} - \begin{bmatrix} P(\sigma) \\ Q(\sigma) \end{bmatrix} u\right). \tag{3.28}$$

It can be seen that the attacked received signal r_s, attack-free input signal u, system matrix $\begin{bmatrix} P(\xi) \\ Q(\xi) \end{bmatrix}$ is given and the check matrix $H(\xi)$ can be computed using the system matrix $\begin{bmatrix} M(\xi) \\ D(\xi) \end{bmatrix}$.

In fact, the signal s is the same as the residual signal in equation (3.26) (this can be proven based on Theorem 3.7). The only difference lies in the system expression (i.e., kernel or ILO image). Moreover, recall equation (3.27), and the parity relations as shown in [10, 29], and is clear that the methods for generating such residual signal s for the ILO image representation are similar to equation (3.27).

For a system given by an ILO image representation, we propose the detection Algorithm 4.

Algorithm 4 Attack detection for system Σ in ILO image representation

1: **procedure** $(M(\xi), D(\xi), P(\xi), Q(\xi), r_s, u, \eta_s)$
▷ Given $M(\xi), D(\xi), P(\xi), Q(\xi), r_s, u$, detect whether η_s is the zero signal.
2: Calculate the check matrix $H(\xi)$ based on equation (3.20).
3: Compute signal $s = H(\sigma)\left(\begin{bmatrix} r_s \\ 0 \end{bmatrix} - \begin{bmatrix} P(\sigma) \\ Q(\sigma) \end{bmatrix} u\right)$.
4: **if** $s = 0$ **then** decide no attack, i.e., $\eta_s = 0$.
5: **else** decide attack occurred, i.e., $\eta_s \neq 0$.
6: **end if**
7: **end procedure**

Theorem 3.10. *Consider a attacked system Σ' under sensor attack given by* (3.16)*. Let r_s, u be the input signals to Algorithm* (4)*, then, the residual signal $s = 0$ if and only if the attack signal $\begin{bmatrix} \eta_s \\ 0 \end{bmatrix} \in \mathcal{B}_{hom}$. Thus Algorithm 4 yields correct result if and only if $\begin{bmatrix} \eta_s \\ 0 \end{bmatrix}$ is detectable or $\eta_s = 0$.*

Proof. Under the detectability assumption, we need to prove the following:

a) when $\eta_s = 0 \Rightarrow s = 0$.

b) when $\eta_s \neq 0 \Rightarrow s \neq 0$.

To prove **a)**, recall (3.16), we have $r_s = y$ and

$$\begin{bmatrix} r_s \\ 0 \end{bmatrix} - \begin{bmatrix} P(\sigma) \\ Q(\sigma) \end{bmatrix} u = \begin{bmatrix} M(\sigma) \\ D(\sigma) \end{bmatrix} l.$$

The residual signal s satisfies

$$s = H(\sigma) \begin{bmatrix} M(\sigma) \\ D(\sigma) \end{bmatrix} l = \begin{bmatrix} 0 & \mathbb{I} \end{bmatrix} \begin{bmatrix} \mathbb{I} \\ 0 \end{bmatrix} l = 0.$$

This completes part **a)** of the proof.

To prove **b)**, assume a detectable sensor attack signal $\eta_s \neq 0$, then

$$\begin{bmatrix} r_s \\ 0 \end{bmatrix} - \begin{bmatrix} P(\sigma) \\ Q(\sigma) \end{bmatrix} u = \begin{bmatrix} M(\sigma) \\ D(\sigma) \end{bmatrix} l + \begin{bmatrix} \eta_s \\ 0 \end{bmatrix},$$

and

$$\begin{aligned} s &= H(\sigma) \begin{bmatrix} M(\sigma) \\ D(\sigma) \end{bmatrix} l - H(\sigma) \begin{bmatrix} \eta_s \\ 0 \end{bmatrix} \\ &= -H(\sigma) \begin{bmatrix} \eta_s \\ 0 \end{bmatrix}. \end{aligned}$$

Now we assume there exists a detectable sensor attack $\eta_s^* \neq 0$ leading to a residual signal $s = 0$; then, we must have $\begin{bmatrix} \eta_s^* \\ 0 \end{bmatrix} \in \ker \{H(\sigma)\}$. It has been proven in Theorem 3.7 that $\ker \{H(\xi)\} = \text{image} \left\{ \begin{bmatrix} M(\xi) \\ D(\xi) \end{bmatrix} \right\}$. This means that for such sensor attack signal η_s^*, there exists a latent signal l^* such that $\begin{bmatrix} M(\sigma) \\ D(\sigma) \end{bmatrix} l^* = \begin{bmatrix} \eta_s^* \\ 0 \end{bmatrix}$. Recalling equation (3.18), it follows that $\begin{bmatrix} \eta_s^* \\ 0 \end{bmatrix} \in \mathcal{B}_{\text{hom}}$ which contradicts the assumption that η_s^* is detectable. This completes the proof. □

Remark: For trivially secure systems in the ILO image representation, the detection algorithm is simply checking whether $s = \begin{bmatrix} r_s \\ 0 \end{bmatrix} - \begin{bmatrix} P(\sigma) \\ Q(\sigma) \end{bmatrix} u$ is zero signal since the attack-free output y is uniquely determinant by the known input u.

3.2.3 Attack Correction for Sensor Attack

We now show how to perform sensor attack correction for systems with known inputs. The objective of attack correction is to return the attack-free sensor output y given the attacked received signal r_s and the known input u.

Before defining our method for attack correction, we need the following definition on a majority vote function which will be used at a later stage in the correction method.

Definition 3.11. *The majority vote function over a set of signals $\{v_1, v_2, \ldots, v_k\}$, denoted by $Maj\{v_1, v_2, \ldots, v_k\}$, is defined to be the most frequently occurring signal in the set of v_j's. If there exists more than one majority signals with equal frequency, then we choose randomly among those majority signals.*

We now present the attack correction algorithms, first for trivially secure systems, and then for general systems that are not trivially secure.

For a trivially secure system in kernel representation where $\delta_s(\Sigma) = m + 1$, we recall that the system is given by Equation (3.10), and thus the following equation regarding the attacked system model holds:

$$r_s - \eta_s = -R_2(\sigma)u.$$

Thereafter, the attack correction can be trivially achieved using the following equation:

$$\eta_s = r_s + R_2(\sigma)u. \tag{3.29}$$

Equivalently, in the ILO image representation, recall that the system is represent as Equation (3.14), attack correction can be achieved by:

$$\eta_s = r_s - P(\sigma)u. \tag{3.30}$$

Theorem 3.11. *If a system under sensor only attack is trivially secure, i.e., $\delta_s(\Sigma) = m + 1$, then any sensor attacks can be corrected.*

We now propose Algorithm 5 regarding the attack correction for general systems.

Algorithm 5 Sensor attack correction algorithm for general system Σ

1: **procedure** $(M(\xi), D(\xi), P(\xi), Q(\xi), r_s, u, \delta_s(\Sigma), \hat{y})$
▷ Given $M(\xi), D(\xi), P(\xi), Q(\xi), r_s, u, \delta_s(\Sigma)$ compute $\hat{y}$.
2: Calculate

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} r_s \\ 0 \end{bmatrix} - \begin{bmatrix} P(\sigma) \\ Q(\sigma) \end{bmatrix} u. \tag{3.31}$$

3: Calculate

$$\hat{l} = \text{Maj}\left\{ \begin{bmatrix} M^{(\mathcal{J},\bullet)}(\sigma) \\ D(\sigma) \end{bmatrix}^{-1} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \right\}, \tag{3.32}$$

where the majority vote is taken over all subsets $\mathcal{J}$ of cardinality $p = m + 1 - \delta_s(\Sigma)$; observers $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) \\ D(\xi) \end{bmatrix}^{-1}$ are defined as in (3.24).
4: $\hat{y} = M(\sigma)\hat{l} + P(\sigma)u$.
5: **return** $\hat{y}$.
6: **end procedure**

Theorem 3.12. *Consider a system Σ with $\delta_s(\Sigma) \leq m$. Assume that $\|\eta_s\| < \delta_s(\Sigma)/2$. Let r_s, u and $\delta_s(\Sigma)$ be inputs to Algorithm 5, then, the estimated output $\hat{y}$ of Algorithm 5 equals y, i.e., the correction algorithm provides the actual attack-free output signal.*

Proof. The proof of the theorem is divided into two parts:

a) We first prove that when the set $\mathcal{J}$ contains only attack-free sensors, then the resulting votes regarding such chosen $\mathcal{J}$ satisfies $\hat{y} = y$. We denote the total number of such correct votes by T.

b) We then prove that the total number of votes for an incorrect signal must be smaller than T.

For **a)**, we first note that given $\|\eta_s\| < \delta_s(\Sigma)/2$, then $\delta_s(\Sigma) - 1 \geq \|\eta_s\|$ must hold. Let $\mathcal{J}$ be a subset of cardinality $p = m + 1 - \delta_s(\Sigma)$ from the set of attack-free sensors. Then,

$$\begin{bmatrix} y^{(\mathcal{J})} + \eta_s^{(\mathcal{J})} \\ 0 \end{bmatrix} = \begin{bmatrix} y^{(\mathcal{J})} \\ 0 \end{bmatrix} = \begin{bmatrix} M^{(\mathcal{J},\bullet)}(\sigma) \\ D(\sigma) \end{bmatrix} l + \begin{bmatrix} P^{(\mathcal{J},\bullet)}(\sigma) \\ Q(\sigma) \end{bmatrix} u,$$

and we have

$$\begin{bmatrix} v_1^{(\mathcal{J})} \\ v_2 \end{bmatrix} = \begin{bmatrix} M^{(\mathcal{J},\bullet)}(\sigma) \\ D(\sigma) \end{bmatrix} l, \tag{3.33}$$

Given $\|\mathcal{J}\|_c = p$, then the left inverse matrix $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) \\ D(\xi) \end{bmatrix}^{-1}$ is well-defined for such chosen $\mathcal{J}$. Multiply such inverse matrix on both sides of equation (3.33) and recall (3.32), the resulting vote $\hat{l}$ for such chosen $\mathcal{J}$ equals the actual latent signal l. There are in total $\binom{m-\|\eta_s\|}{p}$ ways of choosing a subset $\mathcal{J}$ of cardinality p from the set of attack-free sensors. Each choice leads to $\hat{l}$ as the correct signal l; thus, we have $T = \binom{m-\|\eta_s\|}{p}$. This completes part **a)**

For **b)**, we need to prove that the number of the same incorrect votes must be a number smaller than $T = \binom{m-\|\eta_s\|}{p}$ and we prove this by contradiction.

Assume there exists another incorrect vote $l^\star \neq l$ that has at least T votes, then all involved indices $\mathcal{J}$ forms a set, say $\mathcal{F}^\star$. Knowing that if we choose any subset $\mathcal{J} \in \mathcal{F}^\star$ with $\|\mathcal{J}\|_c = p$, the algorithm yields a unique votes. In other words, any p indices from $\mathcal{F}^\star$ leads to the same incorrect signal. In order for the incorrect signal to have at least T votes, the cardinality of $\mathcal{F}^\star$ should be at least $m - \|\eta_s\|$.

Given that $\|\eta_s\| < \delta_s(\Sigma)/2$ and thus $2\|\eta_s\| < \delta_s(\Sigma)$, we have $m - \|\eta_s\| > m - \delta_s(\Sigma) + \|\eta_s\|$. This implies that there are at least $p = m - \delta_s(\Sigma) + 1$ attack-free sensors in $\mathcal{F}^\star$. Since we know that any p attack-free sensors leads to a correct vote l, then $l^\star = l$ must hold which contradicts our assumption that $l^\star \neq l$. Thus we can conclude that there exists no such signal $l^\star \neq l$ that has at least T votes. This completes part **b)** of the proof. □

Notably, Theorem 3.12 provides a sufficient condition that guarantees attack correction. Indeed, when the number of sensors being attacked is $< \delta_s(\Sigma)/2$, Algorithm 5 is guaranteed to generate the attack-free output. When the attack signal is correctable, even when the number of sensors being attacked is $\geq \delta_s(\Sigma)/2$, Algorithm 5 may still provide the correct outcome albeit without any guarantee in those cases.

Comparing our correction Algorithm 5 to the noise-free attack correction method of [13,56], we can see that our algorithm requires fewer observers. In detail, in Section III.A in [13] and Section III.A in [56], the work shows that the number of observers required for a successful attack correction equals to $\binom{m}{m-\delta_s(\Sigma)+1}\binom{m-2\delta_s(\Sigma)+1}{m-2\delta_s(\Sigma)}$ while in our case, the number of observers for a successful attack correction is $\binom{m}{m-\delta_s(\Sigma)+1}$ which is significantly smaller than [13,56].

In the work of [22], a noise-free system under sensor attacks is considered. The attack correction is achieved using an optimisation-based state estimator. In Section 3.C of [22], the l_1/l_r norm was used for the state estimation in order to approximate the ideal l_0 state reconstruction problem, which is NP-hard and highly non-convex. In [22], it is mentioned that based on the result in [9], these two programs are equivalent if and only if the system satisfies

a) all the system matrices in state-space are Gaussian random matrices;

b) the number of the attacked sensors is at the most $0.914\sqrt{m/2}$.

Such assumptions on the system and attack signals limit the performance of the state estimation. The proof of such limitation can be seen on the Numerical Simulation (3.E in [22]), namely, when the number of sensors being attacked is larger than $0.914\sqrt{m/2}$, a successful attack reconstruction is no longer guaranteed. If a system satisfies **a)**, then it is almost certain that the system is trivially secure. In this case, $\delta_s(\Sigma) = m$, meaning our proposed Algorithm 5, is guaranteed to achieve attack correction if the number of sensors being attacked is smaller than $m/2$. Assume $m \geq 2$, then we must have $m/2 \geq 0.914\sqrt{m/2}$. This implies that our algorithm performs better in terms of the upper bound for a guaranteed attack correction.

3.2.4 Sensor Attack Correction Numerical Example

The working of the proposed sensor attack correction method is illustrated using a numerical example.

Consider a discrete-time multi-input multi-output LTI system given by a kernel representation as follows:

$$\begin{bmatrix} \sigma & -1 & 0 \\ 0 & \sigma & -1 \\ -0.5 & 1.5 & \sigma - 1.5 \end{bmatrix} y = \mathbb{I}_3 u.$$

We can see that the system is equivalent to the following system in the state-space form:

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0.5 & -1.5 & 1.5 \end{bmatrix} \quad B = C = \mathbb{I}_3 \quad D = 0.$$

Its Kronecker-Hermite canonical kernel representation as in (3.12) can be computed as:

$$\begin{aligned} R_1(\xi) &= \begin{bmatrix} 1 & 0 & -6\xi^2 + 7\xi - 6 \\ 0 & 1 & -2\xi^2 + 3\xi - 3 \\ 0 & 0 & \xi^3 - 1.5\xi^2 + 1.5\xi - 0.5 \end{bmatrix}, \\ R_2(\xi) &= \begin{bmatrix} -3 & 9 & -6\xi - 2 \\ -1 & 3 & -2\xi \\ 0.5\xi & -1.5\xi + 0.5 & \xi^2 \end{bmatrix}. \end{aligned} \tag{3.34}$$

From equation (3.34), we can see that

$$M(\xi) = \begin{bmatrix} -6\xi^2 + 7\xi - 6 \\ -2\xi^2 + 3\xi - 3 \\ 1 \end{bmatrix}, \quad D(\xi) = \xi^3 - 1.5\xi^2 + 1.5\xi - 0.5.$$

It can be seen that the sensor security index is $\delta_s(\Sigma) = 3$, thus, according to Theorem 3.3, the system Σ can correct any single sensor attack. Polynomial matrices $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) \\ D(\xi) \end{bmatrix}^{-1}$ with $\mathcal{J} = \{1\}, \{2\}, \{3\}$ can be computed as:

$$\begin{bmatrix} M^{(1,\bullet)}(\xi) \\ D(\xi) \end{bmatrix}^{-1} = \begin{bmatrix} \xi^2 & 6\xi + 2 \end{bmatrix}, \quad \begin{bmatrix} M^{(2,\bullet)}(\xi) \\ D(\xi) \end{bmatrix}^{-1} = \begin{bmatrix} \xi & 2 \end{bmatrix}, \quad \begin{bmatrix} M^{(3,\bullet)}(\xi) \\ D(\xi) \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

In the remainder of this section, we present a simulation result based on the system (3.34). The purpose of the simulation is to visualize the working of the proposed majority vote function. Moreover, from the simulation, we visualise the performance of the

proposed algorithm in terms of the amount of delay.

For the sake of simplicity, in this simulation, we consider a unknown additive attack signal on one of the three sensors. Notably, as long as the attack signal is a single sensor attack, the algorithm yields the correct result. Fig. 3.1 shows the three estimated latent signals that are inputs to the majority vote function.

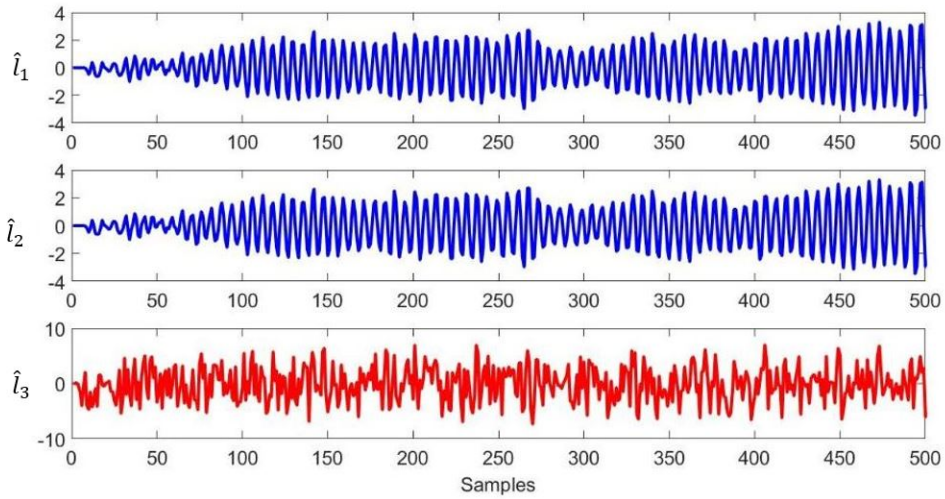

Figure 3.1: Majority vote candidates

As shown by Fig. 3.1, the two blue signals are identical, which determines the majority vote.

Comparing the majority vote output with the actual latent signal l (in this case equals to y_3) which shown in Fig. 3.2, we can see that the majority vote yields a delayed version of the latent signal, which means that our Algorithm 5 indeed reconstructs the correct latent signal and guarantees to achieve attack correction. Fig. 3.3 shows the amount of delay in discrete time for Algorithm 5.

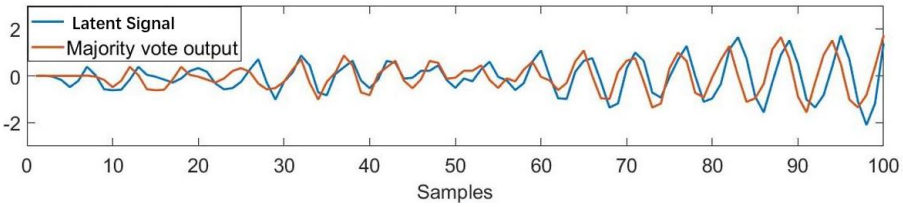

Figure 3.2: Comparison of actual latent signal and majority vote output

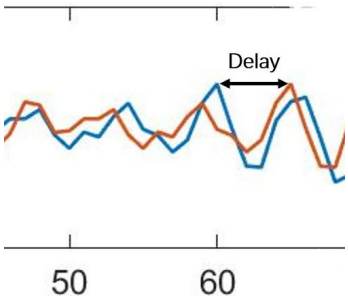

Figure 3.3: Delay for Algorithm 5

As mentioned before, when performing such an attack correction algorithm, the observers are not unique. In fact, for each choice of $\mathcal{J}$, any polynomial matrix that is left unimodularly equivalent with $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) \\ D(\xi) \end{bmatrix}^{-1}$ can generate the same vote. An open research problem followed by this is the process by which we can choose such observers, so that the overall correction delay is minimised. This problem is one of our future research problems.

3.3 Sensor Attack with Prior Sensor Knowledge

At the beginning of this chapter, we have explained that the sensor only attack is the attack scenario being considered in this chapter, and we assume the attacker can potentially attack any of the sensor measurements. However, it is not always the case in many engineering applications. For example, certain sensor measurement signals may have robust encryption/decryption techniques, or certain sensors of a system are physically unreachable by the attacker. As a result, those sensor measurement signals are not accessible by the attacker. This phenomenon can be treated as a specific prior sensor knowledge. In this section, we are going to assume that a specific subset of sensors is never going to be attacked. Under such prior sensor knowledge, we then discuss how to perform attack detection and attack correction.

Intuitively speaking, when we are given such prior sensor knowledge, the system should be more secure against adversarial sensor attacks. Firstly, because now the attacker has limited targets; secondly, because the user can obtain a certain degree of sys-

tem information based on those attack-free sensors. In this section, we will also discuss the benefits of such prior knowledge.

In [30], the concept of security indices α_i is proposed based on certain prior knowledge about sensor attack signals. The difference is that in [30], the set up can be summarised as follows: if the attacker wants to attack sensor i without being detected, what is the total number of other sensors the attacker needs to attack. In our case, we consider the opposite side, namely, if specific sensors are attack-free, how can attack detection and correction be performed? Moreover, we consider prior sensor knowledge for specific sensor sets (rather than individual sensors), which is a more general problem statement.

In this section, we first present the system and attack model based on certain prior sensor knowledge. We then further extend the concept of the security index to systems with prior sensor knowledge. Attack detection and correction methods are subsequently presented. To our knowledge, these results have not been mentioned in previous literature.

3.3.1 Problem Statements and Preliminaries

The attack-free system model being considered in this subsection is the same as in equation (3.2), while the attacked system model is slightly different due to the prior knowledge about the attack signal η_s. In this section, we assume that a subset of sensors with index $\mathcal{I}_s \subsetneq \{1, ..., m\}$ is attack-free. In our scope, an attacked system is given as follows:

$$R_1(\sigma)(r_s - \eta_s) = -R_2(\sigma)u, \text{ where } \eta_s \text{ satisfies } \eta_s^{(\mathcal{I}_s)} = 0. \tag{3.35}$$

Recall that the homogeneous behaviour of the system as in Definition 3.2, the behaviour of interest in this section is a subset of the homogeneous behaviour, which is denoted as $\mathcal{B}_{hom-ps}$, where the subscript 'ps' stands for the prior sensor knowledge and

is defined as follows:

$$\mathcal{B}_{hom-ps} = \left\{ \begin{bmatrix} 0^{(\mathcal{I}_s)} \\ y^{(\bar{\mathcal{I}}_s)} \\ \hline 0 \end{bmatrix} : \mathbb{Z}_+ \to \mathbb{R}^{m+q} \middle| \begin{bmatrix} 0^{(\mathcal{I}_s)} \\ y^{(\bar{\mathcal{I}}_s)} \\ \hline 0 \end{bmatrix} \in \mathcal{B} \right\}. \tag{3.36}$$

The attack detectability and correctability of an attack signal can be defined in a similar way as in Definition 3.6 and 3.7, namely:

A non-zero attack signal $\begin{bmatrix} 0^{(\mathcal{I}_s)} \\ \eta_s^{(\bar{\mathcal{I}}_s)} \\ \hline 0 \end{bmatrix}$ is detectable if $\begin{bmatrix} 0^{(\mathcal{I}_s)} \\ \eta_s^{(\bar{\mathcal{I}}_s)} \\ \hline 0 \end{bmatrix} \notin \mathcal{B}_{hom-ps}$.

A non-zero attack signal $\eta = \begin{bmatrix} 0^{(\mathcal{I}_s)} \\ \eta_s^{(\bar{\mathcal{I}}_s)} \\ \hline 0 \end{bmatrix} \in \mathcal{A}$ is correctable if for all $\eta_s' \neq \eta_s$, the following is satisfied

$$\begin{bmatrix} 0^{(\mathcal{I}_s)} \\ \eta_s'^{(\bar{\mathcal{I}}_s)} \\ \hline 0 \end{bmatrix} \in \mathcal{A} \Rightarrow \begin{bmatrix} 0^{(\mathcal{I}_s)} \\ \eta_s^{(\bar{\mathcal{I}}_s)} \\ \hline 0 \end{bmatrix} - \begin{bmatrix} 0^{(\mathcal{I}_s)} \\ \eta_s'^{(\bar{\mathcal{I}}_s)} \\ \hline 0 \end{bmatrix} \notin \mathcal{B}_{hom-ps}.$$

The new behaviour $\mathcal{B}_{hom-ps}$ is formed by forcing the components with indices $\mathcal{I}_s$ being zero; the reason is that now we are given those sensors that are guaranteed to be attack-free. From the attacker's perspective, in order to remain undetectable, the attacker needs to implement a sensor attack inside $\mathcal{B}_{hom-ps}$.

Further define the security index as follows:

Definition 3.12. *Given that the sensor set $\mathcal{I}_s$ is not accessible by the attacker, the sensor security index of the system Σ is defined as:*

$$\delta_{s-ps}(\Sigma) := \min_{0 \neq \begin{bmatrix} 0^{(\mathcal{I}_s)} \\ y^{(\bar{\mathcal{I}}_s)} \\ \hline 0 \end{bmatrix} \in \mathcal{B}_{hom-ps}} \|y^{(\bar{\mathcal{I}}_s)}\|. \tag{3.37}$$

Further define $\delta_{s-ps}(\Sigma) = \|\bar{\mathcal{I}}_s\| + 1$ if $\mathcal{B}_{hom-ps} = \{0\}$.

Definition 3.13. *Given prior sensor knowledge $\mathcal{I}_s$, we call a system trivially secure if the security index $\delta_{s-ps}(\Sigma) = \|\bar{\mathcal{I}}_s\|_c + 1$.*

Thus, the sensor security index subject to prior sensor knowledge $\mathcal{I}_s$ which is defined in a similar way as the previously defined $\delta_s(\Sigma)$; the difference is that we are now searching inside $\mathcal{B}_{hom-ps}$ instead of $\mathcal{B}_{hom}$.

For systems that are not trivially secure, we propose the following theorem.

Theorem 3.13. *Consider a system Σ satisfying $\delta_{s-ps}(\Sigma) \leq \|\bar{\mathcal{I}}_s\|_c$, then we must have $\delta_{s-ps}(\Sigma) \geq \delta_s(\Sigma)$.*

Proof. Given $\mathcal{B}_{hom-ps} \subseteq \mathcal{B}_{hom}$, the theorem now follows trivially. □

From Theorem 3.13, we can see that in the non-trivially secure case, it is more difficult for the attacker to implement an undetectable/uncorrectable attack when we have certain prior sensor knowledge, simply because $\delta_{s-ps}(\Sigma) \geq \delta_s(\Sigma)$. While in the trivially secure case, any attack signal is detectable and correctable. More detail can be seen in later Section 3.3.2.

In the previous Section 3.2, we have proven that for any system in kernel representation with $R_1(\xi)$ full rank, there exists an image representation as in (3.15) which denotes the same system behaviour. To calculate the security index $\delta_{s-ps}(\Sigma)$, the following theorem is proposed based on the ILO image representation.

Theorem 3.14. *Consider a system Σ in an image representation as in* (3.22) *under sensor attack, further assume prior knowledge: sensors with indices $\mathcal{I}_s$ are attack-free, i.e., $\eta_s^{(\mathcal{I}_s)} = 0$, then the security index $\delta_{s-ps}(\Sigma) = \|\bar{\mathcal{I}}_s\|_c + 1 - k$ where k denotes the smallest integer in $\{0, 1, ..., \|\bar{\mathcal{I}}_s\|_c\}$ such that for any subset $\mathcal{J} \subseteq \bar{\mathcal{I}}_s$ of cardinality k, the $(m + k + \|\mathcal{I}_s\|_c) \times m$ polynomial matrix $\begin{bmatrix} M^{(\mathcal{I}_s,\bullet)}(\xi) \\ M^{(\mathcal{J},\bullet)}(\xi) \\ D(\xi) \end{bmatrix}$ is left unimodular.*

Proof. The proof follows as in Theorem 3.8. □

The detection and correction capability of the system Σ is similar to the previous section (Theorem 3.2 and 3.3). More specifically, we propose the following theorem.

Theorem 3.15. *Consider a system Σ under sensor only attack and given prior sensor knowledge $\mathcal{I}_s$, then an attack signal η is detectable if $\|\eta_s^{(\bar{\mathcal{I}}_s)}\| < \delta_{s-ps}(\Sigma)$. An attack signal η is correctable if $\|\eta_s^{(\bar{\mathcal{I}}_s)}\| < \delta_{s-ps}(\Sigma)/2$. Furthermore, if $\delta_{s-ps} = \|\bar{\mathcal{I}}_s\|_c + 1$, then any η_s can be corrected.*

Proof. The proof follows as in Theorem 3.2 and Theorem 3.3. □

3.3.2 Attack Detection and Correction

The attack detection algorithm for sensor attack, given prior sensor knowledge, is similar to Algorithm 3 and Algorithm 4, by replacing η_s with $\eta_s^{(\bar{\mathcal{I}}_s)}$.

Theorem 3.16. *Detection Algorithm 3 and Algorithm 4 are guaranteed to achieve correct attack detection if the attack signal is detectable or zero. A sufficient condition for a successful attack detection is: $\|\eta_s^{(\bar{\mathcal{I}}_s)}\| < \delta_{s-ps}(\Sigma)$.*

Now that we are in a position to propose a sensor attack correction algorithm for systems with prior knowledge $\mathcal{I}_s$. As a starting point, the system representation being considered is an image representation as in (3.15).

We now propose the attack correction Algorithm 6.

Algorithm 6 Sensor attack correction for systems under prior sensor knowledge

1: **procedure** $(M(\xi), D(\xi), P(\xi), Q(\xi), \mathcal{I}_s, \delta_{s-ps}(\Sigma), r, \hat{y})$
▷ Given $M(\xi)$, $D(\xi)$,$P(\xi)$, $Q(\xi)$, $\mathcal{I}_s$,$\delta_{s-ps}(\Sigma)$ and r, compute $\hat{y}$.
2: Calculate
$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} r_s \\ 0 \end{bmatrix} - \begin{bmatrix} P(\sigma) \\ Q(\sigma) \end{bmatrix} u.$$
3: Calculate
$$\hat{l} = \text{Maj}\left\{ \begin{bmatrix} M^{(\mathcal{I}_s,\bullet)}(\sigma) \\ M^{(\mathcal{J},\bullet)}(\sigma) \\ D(\sigma) \end{bmatrix}^{-1} \begin{bmatrix} v_1^{(\mathcal{I}_s)} \\ v_1^{(\mathcal{J})} \\ v_2 \end{bmatrix} \right\}, \tag{3.38}$$
where the majority vote is taken over all subsets $\mathcal{J}$ of cardinality $\|\bar{\mathcal{I}}_s\|_c + 1 - \delta_{s-ps}(\Sigma)$.
4: Calculate $\hat{y} = M(\sigma)\hat{l} + P(\sigma)u$
5: **return** $\hat{y}$.
6: **end procedure**

Theorem 3.17. *Consider a system Σ given by* (3.1)*; assume sensor subset $\mathcal{I}_s$ is not accessible by the attacker. Denote its security index subject to sensor attack by $\delta_{s-ps}(\Sigma)$. Assume that*

$\eta_s^{(\mathcal{I}_s)} = 0$ and $\|\eta_s^{(\bar{\mathcal{I}}_s)}\| < \delta_{s-ps}(\Sigma)/2$. Let r_s and u be inputs to Algorithm 6, then, the estimated output $\hat{y}$ of Algorithm 6 equals y, i.e., the correction algorithm provides the actual attack-free output signal. Furthermore, when the system is trivially secure, Algorithm 6 provides the actual attack-free output for any attack signals.

Proof. For non-trivially secure case, the proof of this theorem follows as seen in Theorem 3.12. For trivially secure case, we note that the polynomial matrix $\begin{bmatrix} M^{(\mathcal{I}_s,\bullet)}(\xi) \\ D(\xi) \end{bmatrix}^{-1}$ is well-defined. Then there is only one vote for the majority function which involves only attack-free actuators. The proof then follows trivially. □

As compared to the attack correction method without prior sensor knowledge on sensors, we see that the number of observers required in Algorithm 6 is smaller than Algorithm 5. For the trivially secure case, only one observer is needed for system with prior knowledge. For general systems that are not trivially secure, the total number of observers for a system with prior knowledge is $\binom{\|\bar{\mathcal{I}}_s\|_c}{\|\bar{\mathcal{I}}_s\|_c+1-\delta_{s-ps}(\Sigma)}$, while the total number of observers for the same system without prior knowledge is $\binom{m}{m+1-\delta_s(\Sigma)}$. It can be proven that $\binom{\|\bar{\mathcal{I}}_s\|_c}{\|\bar{\mathcal{I}}_s\|_c+1-\delta_{s-ps}(\Sigma)} < \binom{m}{m+1-\delta_s(\Sigma)}$ must hold. This means that with prior knowledge, the computational complexity decreases.

We use the following table to compare and summarise the contents of this chapter.

Table 3.1: Sufficient conditions for sensor only attack detection and correction

<table>
<tr><td colspan="2">Without prior knowledge</td><td colspan="2">With prior knowledge $\mathcal{I}_s$</td></tr>
<tr><td colspan="2">Security index:
$\delta_s = \underset{0 \neq \begin{bmatrix} y \\ 0 \end{bmatrix} \in \mathcal{B}_{hom}}{min} \quad \|y\|$</td><td colspan="2">Security index:
$\delta_{s-ps} = \underset{0 \neq \begin{bmatrix} 0^{(\mathcal{I}_s)} \\ y^{(\bar{\mathcal{I}}_s)} \\ \hline 0 \end{bmatrix} \in \mathcal{B}_{hom-ps}}{min} \quad \|y^{(\bar{\mathcal{I}}_s)}\|$</td></tr>
<tr><td>Trivially secure
$\delta_s = m+1$
Detection
$\|\eta_s\| < \delta_s$
Correction
$\|\eta_s\| < \delta_s$</td><td>Non-trivially secure
$\delta_s < m+1$
Detection
$\|\eta_s\| < \delta_s$
Correction
$\|\eta_s\| < \delta_s/2$</td><td>Trivially Secure
$\delta_{s-ps} = \|\bar{\mathcal{I}}_s\|_c + 1$
Detection
$\|\eta_s^{(\bar{\mathcal{I}}_s)}\| < \delta_{s-ps}$
Correction
$\|\eta_s^{(\bar{\mathcal{I}}_s)}\| < \delta_{s-ps}$</td><td>Non-trivially secure
$\delta_{s-ps} \leq \|\bar{\mathcal{I}}_s\|_c + 1$
Detection
$\|\eta_s^{(\bar{\mathcal{I}}_s)}\| < \delta_{s-ps}$
Correction
$\|\eta_s^{(\bar{\mathcal{I}}_s)}\| < \delta_{s-ps}/2$</td></tr>
</table>

3.4 Recapitulation

In this chapter, we have studied LTI CPS security under sensor only attacks. The main points were:

- A kernel representation was used as the starting point for our research.
- An equivalent ILO image representation was proposed based on the Kronecker-Hermite canonical form of a kernel representation.
- A representation-free system parameter $\delta_s(\Sigma)$ describing the vulnerability of a system against sensor only attacks was proposed. Sufficient conditions with respect to attack detectability and correctability are stated using the notion of $\delta_s(\Sigma)$.
- The notion of $\delta_s(\Sigma)$ also provides a quantitative measurement of a system's sensor redundancy.

- Two universal attack detection methods were proposed (one in kernel representation and the other in ILO image representation)
- An observer-based sensor only attack correction method was proposed.
- Attack detection and correction methods for systems with prior sensor knowledge were discussed.

Chapter 4

Linear Cyber-Physical System Security - Actuator Attacks in the Noise-Free Case

This chapter presents our method for solving the problem of attack detection and attack correction for multi-input multi-output discrete-time linear time-invariant dynamical systems under actuator only attack. In this chapter, the actuator attack signal is modelled as an additive signal without any restriction, with the exception that the number of actuators being attacked is upper bounded by a specific number. In this chapter, we assume that the attack-free inputs of the system are known, and the sensor outputs are also measured. Unlike the case of previous sensor only attack, where the internal latent signal is not corrupted. In the actuator only attack case the latent signal will be corrupted due to the existence of actuator attack. Furthermore, we do not have a direct measurement on either attacked actuator signal or the latent signal.

A kernel representation and an ILO image representation are used to describe the system behaviour. In Section 4.1, we first address the problem of attack detection and attack correction under no prior actuator knowledge, that is, the attacker can potentially attack any actuators. In Section 4.2, we then assume prior actuator knowledge and address attack detection and correction under this assumption. Based on our previous experience with the sensor security index, we propose the notion of 'actuator security index', which is a new addition to the literature. Algorithms for actuator attack detection and attack correction are proposed. Such algorithms are guaranteed to achieve attack detection and attack correction as long as specific upper bounds on the number of attacked actuators

hold.

4.1 Actuator Attack without Prior Actuator knowledge

4.1.1 Problem Statement and Preliminaries

Recall a kernel representation of an attack-free system satisfies

$$\begin{bmatrix} R_1(\sigma) & R_2(\sigma) \end{bmatrix} \begin{bmatrix} y \\ u \end{bmatrix} = 0.$$

Where $y : \mathbb{Z}_+ \to \mathbb{R}^m$ denotes the output signal; $u : \mathbb{Z}_+ \to \mathbb{R}^q$ is the input signal. In this chapter, actuator attack signals η_a are considered. As mentioned in the previous chapter, the square polynomial matrix $R_1(\xi)$ is assumed to have full rank. The resulting attacked system Σ' satisfies the following equation

$$\begin{bmatrix} R_1(\sigma) & R_2(\sigma) \end{bmatrix} \begin{bmatrix} r_s \\ u - \eta_a \end{bmatrix} = 0, \tag{4.1}$$

where r_s denotes the corrupted received signal, $-\eta_a$ is the actuator attack signal. To simplify the later expressions around detectability and correctability, a minus sign is included here. Again, in this chapter, we assume that r_s and u are measured and known by the user.

The ILO image representation plays an important role in this actuator only attack case, because intuitively, the latent signal l is critical when we address the problem of actuator attacks, especially for attack correction.

Recall that the attack-free ILO image representation as in equation (3.15), namely

$$\begin{bmatrix} y \\ 0 \end{bmatrix} = \begin{bmatrix} M(\sigma) & P(\sigma) \\ D(\sigma) & Q(\sigma) \end{bmatrix} \begin{bmatrix} l \\ u \end{bmatrix}.$$

Where $l : \mathbb{Z}_+ \to \mathbb{R}^n$ represents the latent signal; $M(\xi), D(\xi), P(\xi), Q(\xi)$ are polynomial matrices of size $m \times n, n \times n, m \times q, n \times q$ respectively. We assume that the system is

observable, meaning the latent signal l is observable from y for any given u. This implies that the polynomial matrix $\begin{bmatrix} M(\xi) \\ D(\xi) \end{bmatrix}$ is left unimodular (see for example Chapter 5.3 in [46]).

The behaviour $\mathcal{B}$ of the attack-free system is defined as before, see (3.18).

Considering an actuator attack signal η_a, the attacked system Σ' is represented by

$$\begin{bmatrix} r_s \\ 0 \end{bmatrix} = \begin{bmatrix} M(\sigma) & P(\sigma) \\ D(\sigma) & Q(\sigma) \end{bmatrix} \begin{bmatrix} l' \\ u - \eta_a \end{bmatrix}, \tag{4.2}$$

where l' is the corrupted latent signal.

In this chapter, the objective is to detect the existence of the actuator attack η_a, and then reconstruct η_a.

To address the problem regarding actuator attacks, we further define the output-nulling behaviour as follows:

Definition 4.1. *The output-nulling behaviour $\mathcal{B}_N$ of the system Σ is defined as a subset of $\mathcal{B}$ given by*

$$\mathcal{B}_N = \left\{ \begin{bmatrix} 0 \\ u \end{bmatrix} : \mathbb{Z}_+ \to \mathbb{R}^{m+q} \,\middle|\, \begin{bmatrix} 0 \\ u \end{bmatrix} \in \mathcal{B} \right\}. \tag{4.3}$$

Remark: The output-nulling behaviour of the system is a subset of $\mathcal{B}$, as its name suggests, is the collection of all possible input/output pairs whose input u leads to a zero output, i.e. $y = 0$ based on the system dynamics.

In kernel representation, the output-nulling behaviour can be expressed as:

$$\mathcal{B}_N = \left\{ \begin{bmatrix} 0 \\ u \end{bmatrix} : \mathbb{Z}_+ \to \mathbb{R}^{m+q} \,\middle|\, 0 = R_2(\sigma)u \right\}.$$

Equivalently, the output-nulling behaviour can be expressed in terms of the ILO image representation:

$$\mathcal{B}_N := \left\{ \begin{bmatrix} 0 \\ u \end{bmatrix} : \mathbb{Z}_+ \to \mathbb{R}^{m+q} \middle| 0 = \begin{bmatrix} M(\sigma) & P(\sigma) \\ D(\sigma) & Q(\sigma) \end{bmatrix} \begin{bmatrix} l \\ u \end{bmatrix} \text{ for some } l : \mathbb{Z}_+ \to \mathbb{R}^n \right\}. \quad (4.4)$$

Analogous to Definitions 3.6 and 3.7, we now define the attack detectability and attack correctability for actuator attack signals:

Definition 4.2. *A non-zero actuator attack signal* $\begin{bmatrix} 0 \\ \eta_a \end{bmatrix}$ *is detectable if* $\begin{bmatrix} 0 \\ \eta_a \end{bmatrix} \notin \mathcal{B}_N$

The concept of detectable or undetectable actuator may be less intuitive when compared with the sensor case. One may argue that since an undetectable actuator attack does not alter the output, then does such an actuator attack really matter? To explain this argument, we note that if such an undetectable actuator attack does exists, then there are two inputs that give the same output and there is no reason to assume that anything strange happened to the system. However, the actuator attack can alter the internal latent signal of the system without being detected from the input/output knowledge and drive the system into an undesirable performance.

Consider a toy example: a driver controls a vehicle by two types of input signals: wheel angle θ and throttle position p. The wheel angle controls the direction of the vehicle, denoted by d; the throttle position determines the speed of the vehicle v. We further assume that for this system, only v is measured by the speed sensors while the direction d is not being measured. Then, an attacker can attack the wheel angle input θ arbitrarily without being detected. Such an undetectable actuator attack may lead to the vehicle deviating from its original track.

It is now clear as to why having the notion of detectability as in Definition 4.2 is critical. Since this definition indicates what types of actuator attack signals are undetectable, this can potentially damage the system.

Analogous to the previous attack correctability definition, we define an attack set $\mathcal{A}$ that contains all the possible attack signals. We now propose the following definition for actuator attack correctability.

Definition 4.3. *A non-zero attack signal* $\begin{bmatrix} 0 \\ \eta_a \end{bmatrix} \in \mathcal{A}$ *is correctable if for all* $\eta_a' \neq \eta_a$*, the following is satisfied*

$$\begin{bmatrix} 0 \\ \eta_a' \end{bmatrix} \in \mathcal{A} \Rightarrow \begin{bmatrix} 0 \\ \eta_a \end{bmatrix} - \begin{bmatrix} 0 \\ \eta_a' \end{bmatrix} \notin \mathcal{B}_N.$$

The objectives in the following sections are first to determine the conditions for a successful actuator attack detection and then a unique actuator attack correction. We then provide attack detection and correction methods that are guaranteed to produce the correct outcome under certain constraints about the attack signal η_a or the attack set $\mathcal{A}$. In the previous section, a sensor security index was defined in a representation-free context. In order to address the problem of actuator attack, the following representation-free definition of an actuator security index is proposed:

Definition 4.4. *The actuator security index of system* Σ *is defined as*

$$\delta_a(\Sigma) := \min_{0 \neq \begin{bmatrix} 0 \\ u \end{bmatrix} \in \mathcal{B}_N} \|u\|. \tag{4.5}$$

Further define $\delta_a(\Sigma) = q + 1$ *if* $\mathcal{B}_N = \{0\}$.

Definition 4.5. *We call a system trivially secure subject to actuator attack if* $\delta_a(\Sigma) = q + 1$.

Using a kernel representation, the actuator security index can be calculated using only $R_2(\xi)$, as shown in the following theorem.

Theorem 4.1. *The actuator security index satisfies*

$$\delta_a(\Sigma) = k + 1 \tag{4.6}$$

where k is the largest integer in $\{1, ..., q\}$ *such that for any submatrix* $\mathcal{J} \subseteq \{1, ..., q\}$ *of cardinality k, the polynomial matrix* $R_2^{(\bullet, \mathcal{J})}(\xi)$ *is left unimodular.*

Proof. The proof follows as shown in Theorem 3.1, by replacing m with q, R_1 with R_2, $\delta_s(\Sigma)$ with $\delta_a(\Sigma)$ and $\mathcal{B}_{hom}$ with $\mathcal{B}_N$. □

Based on the above theorem, it can be proven that if a system is trivially secure, then the polynomial matrix $R_2(\xi)$ is left unimodular.

The next theorem provides a method for computing the actuator security index $\delta_s(\Sigma)$ based on the ILO image representation (3.15).

Theorem 4.2. *Consider a observable system Σ given by* (3.15)*, then, the actuator security index $\delta_s(\Sigma) = k+1$ where k denotes the largest integer in $\{1,...,q\}$ such that for any subset $\mathcal{J} \subseteq \{1,...,q\}$ of cardinality k, the $(n+m)\times(n+k)$ polynomial matrix $\begin{bmatrix} M(\xi) & P^{(\bullet,\mathcal{J})}(\xi) \\ D(\xi) & Q^{(\bullet,\mathcal{J})}(\xi) \end{bmatrix}$ is left unimodular.*

Proof. We first consider the trivially secure case where $\delta_a(\Sigma) = q+1$. In this case, the only element in $\mathcal{B}_N$ is zero. In this case, the entire polynomial matrix $\begin{bmatrix} M(\xi) & P(\xi) \\ D(\xi) & Q(\xi) \end{bmatrix}$ is left unimodular and thus, we have $k = q$ and $\delta_s(\Sigma) = q+1$, so that $\delta_s(\Sigma) = k+1$.

Now let us consider the case when the system is not trivially secure. Evidently, there exists a subset $\mathcal{J} \subseteq \{1,...,q\}$ of cardinality $\|\mathcal{J}\|_c = k+1$ such that submatrix $\begin{bmatrix} M(\xi) & P^{(\bullet,\mathcal{J})}(\xi) \\ D(\xi) & Q^{(\bullet,\mathcal{J})}(\xi) \end{bmatrix}$ is not left unimodular. Hence, there exists a non-zero signal, say u^* of dimension $k+1$ and a corresponding latent signal l^*, such that

$$\begin{bmatrix} M(\sigma) & P^{(\bullet,\mathcal{J})}(\sigma) \\ D(\sigma) & Q^{(\bullet,\mathcal{J})}(\sigma) \end{bmatrix} \begin{bmatrix} l^* \\ u^* \end{bmatrix} = 0.$$

Now let input signal $u : \mathbb{Z}_+ \to \mathbb{R}^q$ be a signal satisfying $u^{(\mathcal{J})} = u^*$ and $u^{(\bar{\mathcal{J}})} = 0$, clearly $\begin{bmatrix} 0 \\ u \end{bmatrix} \in \mathcal{B}_N$. This implies that $\delta_a(\Sigma) \leq k+1$.

To prove $\delta_a(\Sigma) \geq k+1$, let $\begin{bmatrix} 0 \\ u \end{bmatrix}$ be a signal in $\mathcal{B}_N$ with weight $\|u\| = \delta_a(\Sigma)$. Define a set $\bar{\mathcal{J}}$ of cardinality $\delta_a(\Sigma)$ such that $u^{(\mathcal{J})} = 0$, we then have:

$$\begin{bmatrix} M(\sigma) & P^{(\bullet,\bar{\mathcal{J}})}(\sigma) \\ D(\sigma) & Q^{(\bullet,\bar{\mathcal{J}})}(\sigma) \end{bmatrix} \begin{bmatrix} l \\ u^{\bar{\mathcal{J}}} \end{bmatrix} = 0.$$

Given $u^{(\bar{\mathcal{J}})} \neq 0$, it follows that $\begin{bmatrix} M(\xi) & P^{(\bullet,\bar{\mathcal{J}})}(\xi) \\ D(\xi) & Q^{(\bullet,\bar{\mathcal{J}})}(\xi) \end{bmatrix}$ is not left unimodular. This, in turn, implies that $\delta_a(\Sigma) > k$, and $\delta_a(\Sigma) \geq k+1$. This completes the proof. □

The notion of the trivially secure subject to actuator attacks is equivalent to the concept of strong observability as seen in [57]. In [57], strong observability means that the input signal u can be uniquely determined by the output signal y. In our scope, this is equivalent to polynomial matrix $\begin{bmatrix} M(\xi) & P(\xi) \\ D(\xi) & Q(\xi) \end{bmatrix}$ being left unimodular.

Analogous to Theorem 3.2 and Theorem 3.3, we propose two theorems around the sufficient conditions in order to achieve attack detection and attack correction. Both theorems are stated in terms of the actuator security index $\delta_a(\Sigma)$.

Theorem 4.3. *If the non-zero actuator attack signal η_a satisfies $\|\eta_a\| < \delta_a(\Sigma)$, then $\begin{bmatrix} 0 \\ \eta_a \end{bmatrix}$ is a detectable actuator attack signal.*

Theorem 4.4. *Let the attack set $\mathcal{A}$ be defined as*

$$\mathcal{A} = \left\{ \begin{bmatrix} 0 \\ \eta_a \end{bmatrix} : \mathbb{Z}_+ \to \mathbb{R}^{m+q} \middle| \, \|\eta_a\| < \delta_a(\Sigma)/2 \text{ and } \eta_a \neq 0 \right\}.$$

Then all attacks $\begin{bmatrix} 0 \\ \eta_a \end{bmatrix} \in \mathcal{A}$ are correctable.

The proof of Theorem 4.3 and Theorem 4.4 follow as in Theorem 3.2 and Theorem 3.3.

As seen in the sensor case, $\delta_a(\Sigma)$ is a representation-free concept that characterises the vulnerability of a system against actuator attacks. Analogous to Section 3.2.1, the thresholds $\delta_a(\Sigma)(\delta_a(\Sigma)/2)$ provide sufficient conditios for an actuator attack signal being detectable (correctable), that is, if the number of the actuator being attacked is smaller than $\delta_a(\Sigma)$, then such attack signal must be detectable. If the number of the actuator being attacked is smaller than $\delta_a(\Sigma)/2$, then a unique attack correction is possible.

4.1.2 Attack Detection

The objective of actuator attack detection is to determine whether an actuator attack has occurred. This can be achieved by using detection Algorithm 3 and 4, replacing η_s by η_a. Moreover, we propose the following theorem:

Theorem 4.5. *Detection Algorithm 3 and 4 are guaranteed to achieve correct actuator attack detection if the actuator attack signal is detectable or zero. A sufficient condition for successful attack detection is:* $\|\eta_a\| < \delta_a(\Sigma)$.

Proof. Under the detectability assumption, we need to prove:

a) when $\eta_a = 0 \Rightarrow s = 0$.

b) when $\eta_a \neq 0 \Rightarrow s \neq 0$.

In this proof, we use an ILO image representation (Algorithm 4) as example, but do note that it is also possible to prove this theorem using a kernel representation as shown in Algorithm 3.

To prove **a)**, recall (3.15) and (4.2), we have $r_s = y$ and

$$\begin{bmatrix} r_s \\ 0 \end{bmatrix} - \begin{bmatrix} P(\sigma) \\ Q(\sigma) \end{bmatrix} u = \begin{bmatrix} M(\sigma) \\ D(\sigma) \end{bmatrix} l.$$

The residual signal s satisfies

$$s = H(\sigma) \begin{bmatrix} M(\sigma) \\ D(\sigma) \end{bmatrix} l = \begin{bmatrix} 0 & \mathbb{I} \end{bmatrix} \begin{bmatrix} \mathbb{I} \\ 0 \end{bmatrix} l = 0.$$

This completes part **a)** of the proof.

In order to prove **b)**, assume a detectable actuator attack signal $\eta_a \neq 0$, then

$$\begin{bmatrix} r_s \\ 0 \end{bmatrix} - \begin{bmatrix} P(\sigma) \\ Q(\sigma) \end{bmatrix} u = \begin{bmatrix} M(\sigma) \\ D(\sigma) \end{bmatrix} l' - \begin{bmatrix} P(\sigma) \\ Q(\sigma) \end{bmatrix} \eta_a,$$

and

$$\begin{aligned} s &= H(\sigma)\begin{bmatrix} M(\sigma) \\ D(\sigma) \end{bmatrix} l' - H(\sigma)\begin{bmatrix} P(\sigma) \\ Q(\sigma) \end{bmatrix} \eta_a \\ &= -H(\sigma)\begin{bmatrix} P(\sigma) \\ Q(\sigma) \end{bmatrix} \eta_a = -\begin{bmatrix} 0 & \mathbb{I} \end{bmatrix} U(\sigma)\begin{bmatrix} P(\sigma) \\ Q(\sigma) \end{bmatrix} \eta_a. \end{aligned}$$

Now we assume there exists a detectable actuator attack $\eta_a^* \neq 0$ leading to a residual signal $s = 0$; then, it follows that $\begin{bmatrix} P(\sigma) \\ Q(\sigma) \end{bmatrix} \eta_a^* \in \ker\{H(\sigma)\}$. It has been proven in Theorem 3.7 that $\ker\{H(\xi)\} = \text{image}\left\{\begin{bmatrix} M(\xi) \\ D(\xi) \end{bmatrix}\right\}$. This means that for such actuator attack signal η_a^*, there exists a latent signal l^* such that $\begin{bmatrix} M(\sigma) \\ D(\sigma) \end{bmatrix} l^* = \begin{bmatrix} P(\sigma) \\ Q(\sigma) \end{bmatrix} \eta_a^*$. Then, it follows $\begin{bmatrix} 0 \\ \eta_a^* \end{bmatrix} \in \mathcal{B}_N$ which contradicts the assumption that η_a^* is detectable. This completes the proof. □

4.1.3 Attack Correction

The objective of actuator attack correction is to uniquely return the attack signal η_a given the received signal r_s, the known attack-free input u and the system model. We divide this section into two parts: first, we consider the attack correction algorithms for trivially secure systems, after which we consider general systems that are not trivially secure.

For a trivially secure system in kernel representation where $\delta_a(\Sigma) = q + 1$, without loss of generality, the attacked system can be described as follows:

$$R_1(\sigma) r_s = -(u - \eta_a).$$

Then, the attack correction can be trivially achieved using the following equation:

$$\eta_a = R_1(\sigma) r_s + u.$$

If an ILO image representation is used to describe the system, we are able to recall the following notations around polynomial matrix inverse.

If a polynomial matrix $\begin{bmatrix} M(\xi) & P^{(\bullet,\mathcal{J})}(\xi) \\ D(\xi) & Q^{(\bullet,\mathcal{J})}(\xi) \end{bmatrix}$ is left unimodular, then denote the polynomial left inverse by $\begin{bmatrix} M(\xi) & P^{(\bullet,\mathcal{J})}(\xi) \\ D(\xi) & Q^{(\bullet,\mathcal{J})}(\xi) \end{bmatrix}^{-1}$.

For trivially secure systems in an ILO image representation with $\delta_a(\Sigma) = q+1$, based on Theorem 4.2, we know that the polynomial matrix $\begin{bmatrix} M(\xi) & P(\xi) \\ D(\xi) & Q(\xi) \end{bmatrix}$ is left unimodular. Recall the attacked system equation (4.2), the following equation holds:

$$\begin{bmatrix} l' \\ u - \eta_a \end{bmatrix} = \begin{bmatrix} M(\sigma) & P(\sigma) \\ D(\sigma) & Q(\sigma) \end{bmatrix}^{-1} \begin{bmatrix} r_s \\ 0 \end{bmatrix}.$$

Now, the actuator attack signal can be trivially computed since both r_s and u are known by the user.

To conclude the attack correction for trivially secure systems, we propose the following theorem.

Theorem 4.6. *If a system under actuator only attack is trivially secure, i.e., $\delta_a(\Sigma) = q+1$, then any actuator attack signals can be corrected.*

In the trivially secure case, the correction algorithm is able to correct any actuator attacks, which means that the correctability is bounded by $\delta_a(\Sigma)$ instead of $\delta_a(\Sigma)/2$. This result coincides with the statement in [57]. In Section III, [57], it states that the maximum number of attacked actuators is not inherently restricted by $\lfloor q/2 \rfloor$ and can take value up to q, depending on the specific system under consideration. In our thesis, the specific system structure mentioned in [57] can be recognised as a trivially secure system subject to actuator only attack (or strongly observable using the terminology in [12,57]).

We now propose that the actuator attack correction Algorithm 7 for general systems that are not trivially secure. As mentioned before, the latent signal l is important for actuator only attack; thus, the correction algorithm is proposed based on the ILO image representation.

Algorithm 7 Actuator attack correction for general system Σ

1: **procedure** $(M(\xi), D(\xi), P(\xi), Q(\xi), \delta_a(\Sigma), r_s, u, \hat{\eta}_a)$
▷ Given $M(\xi), D(\xi), P(\xi), Q(\xi), \delta_a(\Sigma), r_s, u$, compute $\hat{\eta}_a$.
2: For each subset $\mathcal{J} \subseteq \{1, ..., q\}$ of cardinality $\|\mathcal{J}\|_c = \delta_a(\Sigma) - 1$, calculates the observer matrix $\begin{bmatrix} M(\xi) & P^{(\bullet,\mathcal{J})}(\xi) \\ D(\xi) & Q^{(\bullet,\mathcal{J})}(\xi) \end{bmatrix}^{-1}$.
3: For each choice of such $\mathcal{J}$, calculate

$$\begin{bmatrix} \hat{l}_{\mathcal{J}} \\ \hat{u}_{\mathcal{J}}^{(\mathcal{J})} \end{bmatrix} = \begin{bmatrix} M(\sigma) & P^{(\bullet,\mathcal{J})}(\sigma) \\ D(\sigma) & Q^{(\bullet,\mathcal{J})}(\sigma) \end{bmatrix}^{-1} \left(\begin{bmatrix} r_s \\ 0 \end{bmatrix} - \begin{bmatrix} P^{(\bullet,\bar{\mathcal{J}})}(\sigma) \\ Q^{(\bullet,\bar{\mathcal{J}})}(\sigma) \end{bmatrix} u^{\bar{\mathcal{J}}} \right). \tag{4.7}$$

4: For each choice of such $\mathcal{J}$, define $\hat{\eta}_{\mathcal{J}} : \mathbb{Z}_+ \to \mathbb{R}^q$ where $\hat{\eta}_{\mathcal{J}}^{(\mathcal{J})} = \hat{u}_{\mathcal{J}}^{(\mathcal{J})}$ and $\hat{\eta}_{\mathcal{J}}^{(\bar{\mathcal{J}})} = 0$.
5: Calculate

$$\begin{bmatrix} \hat{l} \\ \hat{\eta}_a \end{bmatrix} = \text{Maj} \left\{ \begin{bmatrix} \hat{l}_{\mathcal{J}} \\ -\hat{\eta}_{\mathcal{J}} \end{bmatrix} \right\},$$

where the majority vote is taken over all subsets $\mathcal{J} \subseteq \{1, ..., q\}$ with $\|\mathcal{J}\|_c = \delta_a(\Sigma) - 1$.
6: **return** $\hat{\eta}_a$.
7: **end procedure**

Analogous to the sensor attack case, each matrix $\begin{bmatrix} M(\xi) & P^{(\bullet,\mathcal{J})}(\xi) \\ D(\xi) & Q^{(\bullet,\mathcal{J})}(\xi) \end{bmatrix}^{-1}$ can be interpreted as an observer for a chosen subset $\mathcal{J}$. The correction algorithm uses such observers to produce a bank of predictive attack signals, one for each choice of $\mathcal{J}$. Thereafter, the Majority function selects a unique signal among all the resulting signals from the observer outputs.

Theorem 4.7. *Consider a general system under actuator only attack as seen in* (4.2)*. Let the sensor security index $\delta_a(\Sigma)$, received sensor output r_s and the attack-free input signal u be inputs to Algorithm 7. Assume that the actuator attack satisfies $\|\eta_a\| < \delta_a(\Sigma)/2$; then Algorithm 7 uniquely provides the actual actuator attack signal $\hat{\eta}_a = \eta_a$.*

Proof. The proof of the theorem is divided into two parts:

a) We first prove that when the selected subset $\mathcal{J}$ contains all attacked actuators, then the resulting vote regarding such chosen $\mathcal{J}$ satisfies $\hat{\eta}_a = \eta_a$. We denote the total number of such correct votes by p.

we need to prove that the number of votes for an incorrect signal must be a number

smaller

b) Then, we need to prove that the number of votes for an incorrect signal must be a number smaller than p.

To prove **a)**, given $\|\eta_a\| < \delta_a(\Sigma)/2$, then $\delta_a(\Sigma) - 1 \geq \|\eta_a\|$ must hold. Let $\mathcal{J}$ be a subset $\subseteq \{1, ..., q\}$ with $\|\mathcal{J}\|_c = \delta_a(\Sigma) - 1$ that contains all $\|\eta_a\|$ attacked actuators; then, the compliment set $\bar{\mathcal{J}}$ contains only attack-free actuators. For such chosen $\mathcal{J}$, we have the following:

$$\begin{bmatrix} M(\sigma) & P^{(\bullet,\mathcal{J})}(\sigma) \\ D(\sigma) & Q^{(\bullet,\mathcal{J})}(\sigma) \end{bmatrix} \begin{bmatrix} l' \\ u_{\mathcal{J}}^{(\mathcal{J})} \end{bmatrix} = \begin{bmatrix} r_s \\ 0 \end{bmatrix} - \begin{bmatrix} P^{(\bullet,\bar{\mathcal{J}})}(\sigma) \\ Q^{(\bullet,\bar{\mathcal{J}})}(\sigma) \end{bmatrix} u^{\bar{\mathcal{J}}}.$$

Knowing $\|\mathcal{J}\|_c = \delta_a(\Sigma) - 1$, then the inverse matrix $\begin{bmatrix} M(\xi) & P^{(\bullet,\mathcal{J})}(\xi) \\ D(\xi) & Q^{(\bullet,\mathcal{J})}(\xi) \end{bmatrix}^{-1}$ is well-defined.

Multiply both sides by $\begin{bmatrix} M(\xi) & P^{(\bullet,\mathcal{J})}(\xi) \\ D(\xi) & Q^{(\bullet,\mathcal{J})}(\xi) \end{bmatrix}^{-1}$ and recall (4.7), we obtain

$$\begin{bmatrix} \hat{l}_{\mathcal{J}} \\ \hat{u}_{\mathcal{J}}^{(\mathcal{J})} \end{bmatrix} = \begin{bmatrix} l' \\ u^{(\mathcal{J})} - \eta_a^{(\mathcal{J})} \end{bmatrix}.$$

Since $\bar{\mathcal{J}}$ contains only attack-free actuators, then $\eta_a^{(\bar{\mathcal{J}})} = 0$ must hold and thus $\eta^{(\mathcal{J})} = u^{(\mathcal{J})} - \hat{u}_{\mathcal{J}}^{(\mathcal{J})}$, which means that the output of Algorithm 7 satisfies $\hat{\eta}_a = \eta_a$ for such chosen $\mathcal{J}$.

Recall that the choices of such $\mathcal{J}$s are such that $\bar{\mathcal{J}}$ with $\|\bar{\mathcal{J}}\|_c = q + 1 - \delta_a(\Sigma)$ contains only attack-free actuators. There are in total $p = \binom{q - \|\eta_a\|}{q+1-\delta_a(\Sigma)}$ choices of such $\bar{\mathcal{J}}$, i.e., p correct votes for the majority function. This completes part **a)** of the proof.

For part **b)**, we need to prove that the number of votes for an incorrect signal must be a number smaller than $p = \binom{q - \|\eta_a\|}{q+1-\delta_a(\Sigma)}$. We prove this by contradiction.

Assume there exists another incorrect vote say η_a^* that has at least p votes; then, all involving indices $\bar{\mathcal{J}}$ form a set, say $\mathcal{F}^*$. Knowing that if any subset $\bar{\mathcal{J}} \in \mathcal{F}^*$ with $\|\bar{\mathcal{J}}\|_c = q + 1 - \delta_s(\Sigma)$ is chosen, such $\bar{\mathcal{J}}$s will lead to the same incorrect vote. The cardinality of set $\mathcal{F}^*$ should be at least $q - \|\eta_a\|$ in order to have at least p votes.

Given $\|\eta_a\| < \delta_a(\Sigma)/2$ and thus $2\|\eta_a\| < \delta_a(\Sigma)$, we have $q - \|\eta_a\| > q - \delta_a(\Sigma) + \|\eta_a\|$. This implies that there are at least $q - \delta_a(\Sigma) + 1$ attack-free actuators in $\mathcal{F}^*$. Since we know that any $q - \delta_a(\Sigma) + 1$ attack-free actuators lead to a correct vote, then $\eta_a^* = \eta_a$ must hold, which is a contradiction. For this reason, we conclude that there exists no such signal $\eta_a^* \neq \eta_a$ that has at least p votes which completes the proof for part **b)**. □

Analogous to the sensor attack correction Theorem 3.12, it can be seen that Theorem 4.7 provides a sufficient condition that guarantees to perform a unique actuator attack correction. More specifically, when the number of actuators being attacked is $< \delta_a(\Sigma)/2$, Algorithm 7 is guaranteed to generate the correct attack signal.

Despite the fact that several existing literature discussed the issue of actuator attacks ([22, 36] etc.), most of the previous literatures have discussed actuator attacks together with sensor attacks. However, the actuator only case does not attract much attention. In this section, we explore the situation for actuator only attacks (no sensor attack). This setup allows us to understand the essences of the actuator attacks along with the corresponding system properties. The actuator only case provides foundations for a later chapter when both actuator and sensor attacks are presented.

In [57], sensor and actuator attacks are discussed using the notion of (s, r)-sparse strong observability. Potentially, they are looking at actuator only attacks as a special case, namely, by letting the parameter $s = 0$; however, this is not being stated explicitly in [57]. Moreover, comparing our results with [57] regarding such actuator only case, it can be seen that:

- The trivially secure system subject to actuator only attack does not attract much attention in [57]. Their actuator attack correctability is upper bounded by $q/2$ (or even smaller). While in our case, we state explicitly that when the system is strongly observable (or equivalently, $\delta_a(\Sigma) = q + 1$), any actuator attacks can be corrected, and in this case, the upper bound can take value up to q.

- In [57], the upper bound r for actuator only attack correction is presented based on the system representation. Meanwhile in our case, the upper bound is stated in terms of $\delta_a(\Sigma)$, which is a representation-free system parameter.

Moreover, to the best of our knowledge, the proposed detection/correction algorithms, sufficient conditions for detectability/correctability and the concept of actuator security index, are all new addition to the existing body of literature.

4.1.4 Numerical Example

Consider the same numerical example as in Section 3.2.4, namely:

$$\begin{bmatrix} \sigma & -1 & 0 \\ 0 & \sigma & -1 \\ -0.5 & 1.5 & \sigma - 1.5 \end{bmatrix} y = \mathbb{I}_3 u.$$

This time, we consider the actuator only attack. Then, the attacked system satisfies the following equation:

$$\begin{bmatrix} \sigma & -1 & 0 \\ 0 & \sigma & -1 \\ -0.5 & 1.5 & \sigma - 1.5 \end{bmatrix} r = \mathbb{I}_3 (u - \eta_a).$$

It is trivial that $\delta_a(\Sigma) = 4$, i.e., trivially secure, since $R_2(\xi) = \mathbb{I}_3$ is a unimodular matrix. Based on our previous discussion any actuator attack can be corrected trivially in this case.

Now, let us consider a dynamical system under actuator only attack that is not trivially secure.

Consider a system that satisfies the following equation:

$$\begin{bmatrix} \sigma - 1 & 0 & 0 \\ 0 & \sigma - 1 & 0 \\ 0 & 0 & \sigma - 1 \end{bmatrix} y = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 2 \end{bmatrix} u.$$

The system is not trivially secure since the matrix $R_2(\xi) = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 2 \end{bmatrix}$ is not left unimod-

ular. Based on Theorem 4.1, it can be seen that $\delta_a(\Sigma) = 3$. From another point of view, there exists an input signal with weight $= 3$ that lies in the output nulling behaviour, i.e., $u(t) = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$ for all $t \in \mathbb{Z}_+$. We can see that the system is equivalent to the following system in the state-space form:

$$A = \mathbb{I}_3, B = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 2 \end{bmatrix}, C = \mathbb{I}_3 \text{ and } D = 0.$$

We now express the system under actuator only attack in the following ILO image representation:

$$\begin{bmatrix} r \\ 0 \end{bmatrix} = \left[\begin{array}{ccc|ccc} & \mathbb{I}_3 & & & 0_{3\times 3} & \\ \hline \sigma-1 & 0 & 0 & 0 & -1 & -1 \\ 0 & \sigma-1 & 0 & -1 & 0 & -1 \\ 0 & 0 & \sigma-1 & -1 & -1 & -2 \end{array}\right] \begin{bmatrix} l \\ u-\eta_a \end{bmatrix}.$$

Then, the three observers can be calculated as

$$\begin{bmatrix} M(\xi) & P^{(\bullet,\mathcal{J})}(\xi) \\ D(\xi) & Q^{(\bullet,\mathcal{J})}(\xi) \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \text{ for } \mathcal{J} = \{1,2\}$$

$$\begin{bmatrix} M(\xi) & P^{(\bullet,\mathcal{J})}(\xi) \\ D(\xi) & Q^{(\bullet,\mathcal{J})}(\xi) \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \end{bmatrix} \text{ for } \mathcal{J} = \{1,3\}$$

$$\begin{bmatrix} M(\xi) & P^{(\bullet,\mathcal{J})}(\xi) \\ D(\xi) & Q^{(\bullet,\mathcal{J})}(\xi) \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \end{bmatrix} \text{ for } \mathcal{J} = \{2,3\}.$$

4.2 Actuator Attack with Prior Actuator Knowledge

In addition to the 'actuator only attack' knowledge, in this section, we further assume that a subset of actuators is not accessible by the attacker. We denote such an assumption as a specific prior actuator knowledge.

In many engineering applications, some actuator input signals are generated by the user and directly applied to the targeted system. For example, in a combustion vehicle, the throttle position controlled by the driver will directly influence air intake. Thus, it is generally impossible for the attacker to gain access to this actuator signal. For these reasons, in many engineering applications, it is necessary to take such prior actuator knowledge into account, and that is the main motivation for us to present this topic.

In this section, we first present the system and attack model based on the prior actuator knowledge. We then define the security index subject to such prior actuator knowledge. Attack detection and correction methods are presented based on the proposed security index. To the best of our knowledge, these results are new additions to extant literature.

4.2.1 Problem Statements and Preliminaries

The attack-free system model that is considered in this subsection is an ILO image representation, which is the same as seen in equation (3.15). In this subsection, prior actuator knowledge is considered where we assume the subset of actuators with index

$\mathcal{I}_a \subsetneq \{1, ..., q\}$ is attack-free. In our scope, an attacked system is given as follows:

$$\begin{bmatrix} r_s \\ 0 \end{bmatrix} = \begin{bmatrix} M(\sigma) & P(\sigma) \\ D(\sigma) & Q(\sigma) \end{bmatrix} \begin{bmatrix} l' \\ u - \eta_a \end{bmatrix}, \text{ where } \eta_a \text{ satisfies } \eta_a^{(\mathcal{I}_a)} = 0. \tag{4.8}$$

The actuator attack signal η_a can therefore be expressed as $\eta_a = \begin{bmatrix} 0^{(\mathcal{I}_a)} \\ \eta_a^{(\bar{\mathcal{I}}_a)} \end{bmatrix}$.

Recall that the output-nulling behaviour of the system as in Definition 4.1 and equation (4.4), since we are given prior actuator knowledge; then, the behaviour of interest in this section is a subspace of the output-nulling behaviour, denoted as $\mathcal{B}_{N-pa}$, where the subscript 'pa' stands for prior actuator knowledge. $\mathcal{B}_{N-pa}$ is defined as follows:

$$\mathcal{B}_{N-pa} := \left\{ \begin{bmatrix} 0 \\ \hline 0^{(\mathcal{I}_a)} \\ u^{(\bar{\mathcal{I}}_a)} \end{bmatrix} : \mathbb{Z}_+ \to \mathbb{R}^{m+q} \,\middle|\, \begin{bmatrix} 0 \\ \hline 0^{(\mathcal{I}_a)} \\ u^{(\bar{\mathcal{I}}_a)} \end{bmatrix} \in \mathcal{B}_N \right\}. \tag{4.9}$$

The attack detectability and correctability of an attack signal can be defined in a similar way as shown in Definition 4.2 and Definition 4.3, namely:

A non-zero attack signal $\begin{bmatrix} 0 \\ \hline 0^{(\mathcal{I}_a)} \\ \eta_a^{(\bar{\mathcal{I}}_a)} \end{bmatrix}$ is detectable if $\begin{bmatrix} 0 \\ \hline 0^{(\mathcal{I}_a)} \\ \eta_a^{(\bar{\mathcal{I}}_a)} \end{bmatrix} \notin \mathcal{B}_{N-pa}$.

A non-zero attack signal $\eta_a = \begin{bmatrix} 0 \\ \hline 0^{(\mathcal{I}_a)} \\ \eta_a^{(\bar{\mathcal{I}}_a)} \end{bmatrix} \in \mathcal{A}$ is correctable if for all $\begin{bmatrix} 0 \\ \hline 0^{(\mathcal{I}_a)} \\ \eta_a'^{(\bar{\mathcal{I}}_a)} \end{bmatrix} \neq \begin{bmatrix} 0 \\ \hline 0^{(\mathcal{I}_a)} \\ \eta_a^{(\bar{\mathcal{I}}_a)} \end{bmatrix}$, the following relation holds:

$$\begin{bmatrix} 0 \\ \hline 0^{(\mathcal{I}_a)} \\ \eta_a'^{(\bar{\mathcal{I}}_a)} \end{bmatrix} \in \mathcal{A} \Rightarrow \begin{bmatrix} 0 \\ \hline 0^{(\mathcal{I}_a)} \\ \eta_a^{(\bar{\mathcal{I}}_a)} \end{bmatrix} - \begin{bmatrix} 0 \\ \hline 0^{(\mathcal{I}_a)} \\ \eta_a'^{(\bar{\mathcal{I}}_a)} \end{bmatrix} \notin \mathcal{B}_{N-pa}.$$

Further, define the actuator security index with prior actuator knowledge as follows:

Definition 4.6. *Given that the actuator set $\mathcal{I}_a$ is not accessible by the attacker, the actuator*

security index of the system Σ *is defined as:*

$$\delta_{a-pa}(\Sigma) = \min_{0 \neq \begin{bmatrix} 0 \\ \hline 0^{(\mathcal{I}_a)} \\ u^{(\bar{\mathcal{I}}_a)} \end{bmatrix} \in \mathcal{B}_{N-pa}} \|u^{(\bar{\mathcal{I}}_a)}\|. \tag{4.10}$$

Further define $\delta_{a-pa}(\Sigma) = \|\bar{\mathcal{I}}_a\|_c + 1$ *if* $\mathcal{B}_{N-pa} = \{0\}$.

Definition 4.7. *We call a system trivially secure if the actuator security index* $\delta_{a-pa}(\Sigma) = \|\bar{\mathcal{I}}_a\|_c + 1$.

Analogous to section 3.3.1, we propose the following theorem:

Theorem 4.8. *Consider a system* Σ *satisfies* $\delta_{a-pa}(\Sigma) \leq \|\bar{\mathcal{I}}_a\|_c$*, then we must have* $\delta_{a-pa}(\Sigma) \geq \delta_a(\Sigma)$.

Proof. Given $\mathcal{B}_{N-pa} \subset \mathcal{B}_N$, the proof now follows trivially. □

Thus, Theorem 4.8 implies that it is more challenging for the attacker to implement an undetectable/uncorrectable actuator attack in the sense that more actuators need to be attacked if we are given such prior actuator knowledge, since $\delta_{a-pa}(\Sigma) \geq \delta_a(\Sigma)$.

To calculate $\delta_{a-pa}(\Sigma)$, we propose the following theorem.

Theorem 4.9. *Consider the attacked system model as in* (4.8)*, then the security index* $\delta_{a-pa}(\Sigma) = k + 1$ *where* k *is the largest integer in* $\{0, 1, ..., \|\bar{\mathcal{I}}_a\|_c\}$ *such that for any subset* $\mathcal{J} \subseteq \bar{\mathcal{I}}_a$ *of cardinality* k*, the polynomial matrix* $\begin{bmatrix} M(\xi) & P^{(\bullet,\mathcal{I}_a)}(\xi) & P^{(\bullet,\mathcal{J})}(\xi) \\ D(\xi) & Q^{(\bullet,\mathcal{I}_a)}(\xi) & Q^{(\bullet,\mathcal{J})}(\xi) \end{bmatrix}$ *is left unimodular.*

Proof. The proof follows as in Theorem 4.2. □

Sufficient conditions for the detection and correction capability for actuator only attack with prior actuator knowledge $\mathcal{I}_s$ is given in the following theorem.

Theorem 4.10. *Consider a system* Σ *under actuator only attack and given prior actuator knowledge* $\mathcal{I}_a$*. Then, an attack signal* η *is detectable if* $\|\eta_a^{(\bar{\mathcal{I}}_a)}\| < \delta_{a-pa}(\Sigma)$*. An attack signal* η *is correctable if* $\|\eta_a^{(\bar{\mathcal{I}}_a)}\| < \delta_{a-pa}(\Sigma)/2$*. Furthermore, if* $\delta_{a-pa} = \|\bar{\mathcal{I}}_a\|_c + 1$*, then any* η_a *can be corrected.*

Proof. The proof follows as shown in Theorem 3.2 and Theorem 3.3. □

4.2.2 Attack Detection and Correction

The attack detection algorithm for actuator attack given prior actuator knowledge is identical with Algorithm 3 and Algorithm 4, replacing η_s by $\eta_a^{(\bar{\mathcal{I}}_a)}$. Moreover, we propose the following theorem.

Theorem 4.11. *Detection Algorithm 3 and 4 are guaranteed to achieve correct actuator attack detection if the actuator attack signal is detectable or zero. A sufficient condition for a successful attack detection is:* $\|\eta_a^{(\bar{\mathcal{I}}_a)}\| < \delta_{a-pa}(\Sigma)$.

Meanwhile we propose attack correction Algorithm 8.

Algorithm 8 Actuator attack correction for systems under prior Actuator knowledge

1: **procedure** $(M(\xi), D(\xi), P(\xi), Q(\xi), \mathcal{I}_a, \delta_{a-pa}(\Sigma), r_s, u, \hat{\eta}_a^{(\bar{\mathcal{I}}_a)})$
▷ Given $M(\xi), D(\xi), P(\xi), Q(\xi), \mathcal{I}_a, \delta_{a-pa}(\Sigma), r_s, u$, compute $\hat{\eta}_a^{(\bar{\mathcal{I}}_a)}$.
2: For each subset $\mathcal{J} \subseteq \bar{\mathcal{I}}_a$ of cardinality $\|\mathcal{J}\|_c = \delta_{a-pa}(\Sigma) - 1$, calculates the observer matrix $\begin{bmatrix} M(\xi) & P^{(\bullet,\mathcal{I}_a)}(\xi) & P^{(\bullet,\mathcal{J})}(\xi) \\ D(\xi) & Q^{(\bullet,\mathcal{I}_a)}(\xi) & Q^{(\bullet,\mathcal{J})}(\xi) \end{bmatrix}^{-1}$.
3: For each choice of such $\mathcal{J}$, calculate

$$\begin{bmatrix} \hat{l}_{\mathcal{J}} \\ \hat{u}_{\mathcal{J}}^{(\mathcal{J})} \end{bmatrix} = \begin{bmatrix} M(\sigma) & P^{(\bullet,\mathcal{I}_a)}(\sigma) & P^{(\bullet,\mathcal{J})}(\sigma) \\ D(\sigma) & Q^{(\bullet,\mathcal{I}_a)}(\sigma) & Q^{(\bullet,\mathcal{J})}(\sigma) \end{bmatrix}^{-1} \left(\begin{bmatrix} r_s \\ 0 \end{bmatrix} - \begin{bmatrix} P^{(\bullet,\bar{\mathcal{J}})}(\sigma) \\ Q^{(\bullet,\bar{\mathcal{J}})}(\sigma) \end{bmatrix} u^{\bar{\mathcal{J}}} \right), \tag{4.11}$$

where the $\bar{\mathcal{J}}$ in the above equation denotes the complement set of $\mathcal{J}$ with respect to set $\bar{\mathcal{I}}_a$.
4: For each choice of such $\mathcal{J}$, define $\hat{\eta}_{\mathcal{J}} : \mathbb{Z}_+ \to \mathbb{R}^q$ where $\hat{\eta}_{\mathcal{J}}^{(\mathcal{J})} = \hat{u}_{\mathcal{J}}^{(\mathcal{J})}$, $\eta_{\mathcal{J}}^{(\mathcal{I}_a)} = 0$ and $\hat{\eta}_{\mathcal{J}}^{(\bar{\mathcal{J}})} = 0$.
5: Calculate

$$\begin{bmatrix} \hat{l} \\ \hat{\eta}_a \end{bmatrix} = \text{Maj} \left\{ \begin{bmatrix} \hat{l}_{\mathcal{J}} \\ -\hat{\eta}_{\mathcal{J}} \end{bmatrix} \right\},$$

where the majority vote is taken over all subsets $\mathcal{J} \subseteq \bar{\mathcal{I}}_a$ with $\|\mathcal{J}\|_c = \delta_{a-pa}(\Sigma) - 1$.
6: **return** $\hat{\eta}_a$.
7: **end procedure**

Theorem 4.12. *Consider a system* Σ *under actuator only attack with prior actuator knowledge as seen in equation* (4.8)*. Let* r_s*,* u*,* $\mathcal{I}_a$ *and the security index* $\delta_{a-pa}(\Sigma)$ *be the inputs to Algorithm 8.*

Assume the actuator attack signal satisfies $\|\eta_a^{(\bar{\mathcal{I}}_a)}\| < \delta_{a-pa}(\Sigma)/2$*, then the Algorithm 8 provides the correct attack signal, i.e.* $\hat{\eta}_a = \begin{bmatrix} 0^{(\mathcal{I}_a)} \\ \hat{\eta}_a^{(\bar{\mathcal{I}}_a)} \end{bmatrix} = \begin{bmatrix} 0^{(\mathcal{I}_a)} \\ \eta_a^{(\bar{\mathcal{I}}_a)} \end{bmatrix}$*. Furthermore, when the system is trivially secure, Algorithm 8 provides the actual signal for any attack signal.*

Proof. The proof follows as shown in Theorem 4.7. □

We use the following table to compare and summarise the contents of this chapter.

Table 4.1: Sufficient conditions for actuator only attack detection and correction

Without prior knowledge		With prior knowledge $\mathcal{I}_s$	
Security index: $\delta_a(\Sigma) := \min_{0 \neq \begin{bmatrix} 0 \\ u \end{bmatrix} \in \mathcal{B}_N} \|u\|$		Security index: $\delta_{a-pa}(\Sigma) = \min_{0 \neq \left[\begin{array}{c} 0 \\ \hline 0^{(\mathcal{I}_a)} \\ u^{(\bar{\mathcal{I}}_a)} \end{array}\right] \in \mathcal{B}_{N-pa}} \|u^{(\bar{\mathcal{I}}_a)}\|$	
Trivially secure	Non-trivially secure	Trivially Secure	Non-trivially secure
$\delta_a = q+1$	$\delta_a < q+1$	$\delta_{a-pa} = \|\bar{\mathcal{I}}_a\|_c + 1$	$\delta_{a-pa} \leq \|\bar{\mathcal{I}}_a\|_c + 1$
Detection	Detection	Detection	Detection
$\|\eta_a\| < \delta_a$	$\|\eta_a\| < \delta_a$	$\|\eta_a^{(\bar{\mathcal{I}}_a)}\| < \delta_{a-pa}$	$\|\eta_a^{(\bar{\mathcal{I}}_a)}\| < \delta_{a-pa}$
Correction	Correction	Correction	Correction
$\|\eta_a\| < \delta_a$	$\|\eta_a\| < \delta_a/2$	$\|\eta_a^{(\bar{\mathcal{I}}_a)}\| < \delta_{a-pa}$	$\|\eta_a^{(\bar{\mathcal{I}}_a)}\| < \delta_{a-pa}/2$

4.3 Recapitulation

In this chapter, we have studied LTI CPS security under actuator only attacks. The main points discussed were:

- An undetectable actuator attack is dangerous to the targeted system because the attacker can alter the latent signal without changing the received outputs.
- We extended the concept of security index in this actuator only attack case, namely, actuator security index $\delta_a(\Sigma)$.

- Detectability and correctability for actuator only attacks were stated using the notion of $\delta_a(\Sigma)$.
- Trivially secure system subject to actuator only attack is equivalent to strong observability of a system.
- An observer-based actuator only attack correction method was proposed.
- Attack detection and correction methods for systems with prior actuator knowledge were discussed.

Chapter 5

Linear Cyber-Physical System Security - Sensor and Actuator Attacks in the Noise-Free Case

This chapter presents our methods for solving the problem of attack detection and attack correction for multi-input multi-output discrete-time linear time-invariant dynamical system under both sensor and actuator attacks.

The attack scenario considered in this chapter is more complex as compared to previous chapters since both sensor and actuator can be attacked. In [57], the problem of state estimation under both actuator and sensor attacks is discussed; even so there remain many open questions. One particular open research question draws our attention. In [57], the authors observe the fact that 'the maximum number of attacked actuators is not inherently restricted by $\lfloor q/2 \rfloor$ and can take value up to q', but a detailed explanation is absent [57].

We start with the question: is it possible to achieve guaranteed sensor and actuator attacks correction when **all** actuators are attacked together with some (but not all) sensors? Also, can we apply our developed techniques regarding attack detection/correction for this more general attack model? In order to achieve our aims, it turns out that the property of strong observability is essential. The notion of strong observability can be interpreted as: the input signal u can be uniquely determined from knowledge of the system dynamics and the output signal y. Without this property, there may exist infinitely many different inputs, all leading to the same received output. Put differently, without strong observability, it is impossible to achieve correction in the case of an attack. For this rea-

son, in this chapter, we assume that the systems (or certain sub-systems) possess strong observability.

In Section 5.1, we address the problem of attack detection and attack correction under no prior sensor and/or actuator knowledge. An important feature for this attack scenario is that we do not have any prior knowledge about the attack signal, that is, both the sensor attack signal η_s and the actuator attack signal η_a are potentially non-zero; moreover, the attacker can potentially access any sensors/actuators.

Based on our previous knowledge around security index, a new concept of sensor and actuator security index is proposed in this section, to address the problem of detectability of the sensor and actuator attacks. Unlike the case of previous chapters where the correctability can also be stated in terms of the same security index a new concept of 'correction index' is proposed for attack correction in this section. The correction index is generally not equal to the sensor and actuator security index.

The objective for attack detection is the same as before, namely, to determine whether an attack has occurred. However, as we mentioned before, for attack correction, we are aiming to correct for any actuator attacks and specific upper-bounded sensor attacks. Since the strong observability is assumed, the sensor signal plays a more significant role in attack correction than the actuator signal. Once the sensor attack is corrected, it is then possible for us to correct any actuator attacks due to the strong observability of the system.

Knowing that both sensor and actuator are potentially attacked, in Section 5.2, we further assume certain prior sensor knowledge; namely, we assume that a subset of sensors is not accessible by the attacker. In Section 5.3, the prior knowledge is on the actuators. In both sections, we propose attack detection/correction methods. Finally, we briefly discuss the problem of attack detection and correction for sensor and actuator attacks with prior knowledge on both sensor and actuator in Section 5.4.

To the best our knowledge, the correction index, sensor and actuator security index and the proposed detection/correction algorithms for different attack models presented in this chapter are new additions to extant literature.

5.1 Attack Detection and Correction for Actuator and Sensor Attack

5.1.1 Preliminaries and Detectability

Recall the attack-free system in its kernel representation, we have

$$\begin{bmatrix} R_1(\sigma) & R_2(\sigma) \end{bmatrix} \begin{bmatrix} y \\ u \end{bmatrix} = 0, \tag{5.1}$$

where the output signal $y : \mathbb{Z}_+ \to \mathbb{R}^m$, input signal $u : \mathbb{Z}_+ \to \mathbb{R}^q$. The system dynamics under both actuator and sensor attack follows:

$$\begin{bmatrix} R_1(\sigma) & R_2(\sigma) \end{bmatrix} \begin{bmatrix} r_s - \eta_s \\ u - \eta_a \end{bmatrix} = 0. \tag{5.2}$$

As seen in previous chapters, signal r_s is the received sensor signal. Signal η_s and η_a are the sensor and actuator attacks respectively. As in previous chapters, signal r_s and u are measured and known by the user.

As for the equivalent ILO image representation, consider a observable linear system Σ, whose latent signal l is observable from y for any given u. The dynamics of the system are given by an ILO image representation as in (3.15), namely

$$\begin{bmatrix} y \\ 0 \end{bmatrix} = \begin{bmatrix} M(\sigma) & P(\sigma) \\ D(\sigma) & Q(\sigma) \end{bmatrix} \begin{bmatrix} l \\ u \end{bmatrix},$$

where the latent signal $l : \mathbb{Z}_+ \to \mathbb{R}^n$. The behaviour of Σ is defined as in equation (3.18). For a system under actuator and sensor attacks, we propose the following attacked system representation:

$$\begin{bmatrix} r_s - \eta_s \\ 0 \end{bmatrix} = \begin{bmatrix} M(\sigma) & P(\sigma) \\ D(\sigma) & Q(\sigma) \end{bmatrix} \begin{bmatrix} l' \\ u - \eta_a \end{bmatrix}. \tag{5.3}$$

The attack signals result in a corrupted latent signal, which we denote as l'. Moreover,

we denote the attack signal η as follows:

$$\eta := \begin{bmatrix} \eta_s \\ \eta_a \end{bmatrix}. \tag{5.4}$$

The behaviour of interest in this chapter is different in comparison to previous chapters. Since we are now considering both sensor and actuator attacks, the entire behaviour $\mathcal{B}$ that contains all possible input/output pairs is now key.

Recall the definition of system behaviour $\mathcal{B}$ as in equation (3.3) and (3.18). We now propose the following definition regarding attack detectability.

Definition 5.1. *A non-zero attack signal* $\eta = \begin{bmatrix} \eta_s \\ \eta_a \end{bmatrix}$ *is detectable if* $\eta \notin \mathcal{B}$.

Analogous to previous chapters, the sensor and actuator security index $\delta_{sa}(\Sigma)$ is defined as follows, where the subscript 'sa' stands for sensor and actuator attacks:

Definition 5.2. *The sensor and actuator security index* $\delta_{sa}(\Sigma)$ *of system* Σ *is defined as:*

$$\delta_{sa}(\Sigma) := \min_{0 \neq \begin{bmatrix} y \\ u \end{bmatrix} \in \mathcal{B}} \left\| \begin{bmatrix} y \\ u \end{bmatrix} \right\|. \tag{5.5}$$

We note that the concept of trivially secure is not applicable in this setup, that is, $\mathcal{B} \neq \{0\}$ must hold, since the input signal u can be arbitrary.

In order to calculate the sensor and actuator security index, we choose to use the kernel representation, as in equation (5.1), and further denote $R(\xi) = \begin{bmatrix} R_1(\xi) & R_2(\xi) \end{bmatrix}$ as in previous chapters.

Theorem 5.1. *The sensor and actuator security index can be computed as:* $\delta_{sa}(\Sigma) = k+1$ *where* k *is the largest integer in* $\{1, ..., m\}$ *such that any submatrix* $R^{(\bullet, \mathcal{J})}(\xi)$ *with* $\mathcal{J} \subseteq \{1, ..., m+q\}$ *and* $\|\mathcal{J}\|_c = k$ *is left unimodular.*

Proof. The proof of the theorem follows as in Theorem 3.8. □

Remark: It is also possible to calculate $\delta_{sa}(\Sigma)$ if we choose to use the ILO image representation as the starting point. Since the ILO image representation is equivalent to the kernel representation using the check matrix $H(\xi)$, i.e., $R(\xi) = H(\xi)\begin{bmatrix} \mathbb{I} & -P(\xi) \\ 0 & -Q(\xi) \end{bmatrix}$.

A sufficient condition for attack detectability can be stated similar to what is seen in Theorem 4.3 as follows.

Theorem 5.2. *If a non-zero attack signal* $\eta = \begin{bmatrix} \eta_s \\ \eta_a \end{bmatrix}$ *satisfies* $\|\eta\| < \delta_{sa}(\Sigma)$*, then* η *is a detectable attack signal.*

Algorithm 3 and 4 can be used as the attack detection algorithms for a system under both actuator and sensor attacks, replacing η_s by η. Moreover, we propose the following theorem.

Theorem 5.3. *Detection Algorithm 3 and 4 are guaranteed to achieve correct actuator and sensor attack detection if the attack signal* η *is detectable or zero.*

5.1.2 Attack Correction for Actuator and Sensor Attack

Recall that the objective of the attack correction is to generate the actual attack signal $\eta = \begin{bmatrix} \eta_s \\ \eta_a \end{bmatrix}$ uniquely. As a first step towards proposing the attack correction algorithm, we introduce a definition that will be used in later discussions, namely, the 'correction index'.

Definition 5.3. *The correction index* $\delta_c(\Sigma)$ *of system* Σ *is defined as:*

$$\delta_c(\Sigma) := \min_{0 \neq \begin{bmatrix} y \\ u \end{bmatrix} \in \mathcal{B}} \|y\|. \tag{5.6}$$

The definition of the correction index $\delta_c(\Sigma)$ is similar in some perspectives to the previously defined security index $\delta_{sa}(\Sigma)$. Both these definitions requires a search within the entire behaviour. However, the major difference lies in the argument of the definitions,

for security index, minimisation objective is $\begin{bmatrix} y \\ u \end{bmatrix}$ while the correction index only considers the sub-signal y. Moreover, we propose the following lemma:

Lemma 5.1. *Consider a system Σ under actuator and sensor attacks, then the following equation must hold:*

$$\delta_c(\Sigma) \leq \delta_{sa}(\Sigma).$$

Proof. Recall equation (5.5) and (5.6), we see that the minimisation objective of $\delta_c(\Sigma)$ is y, which is a sub-signal as compared to the objective of $\delta_{sa}(\Sigma)$, which is $\begin{bmatrix} y \\ u \end{bmatrix}$. The proof now follows trivially. □

The following theorem provides a calculation method for computing the correction index $\delta_c(\Sigma)$. Since the attack take place on both sensors and actuators and thus a latent signal plays a vital role. For this reasoning, an ILO image representation is considered here.

Theorem 5.4. *The correction index $\delta_c(\Sigma)$ for system Σ satisfies $\delta_c(\Sigma) = m + 1 - k$ where k denotes the smallest integer in $\{1, ..., m\}$ such that for any subset $\mathcal{J} \subseteq \{1, ..., m\}$ of cardinality $\|\mathcal{J}\|_c = k$, the $(k+n) \times (q+n)$ polynomial matrix*

$$\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) & P^{(\mathcal{J},\bullet)}(\xi) \\ D(\xi) & Q(\xi) \end{bmatrix}$$

is left unimodular.

Proof. Clearly there exists a subset $\mathcal{J} \subseteq \{1, \ldots, m\}$ of cardinality $k-1$ such that

$$\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) & P^{(\mathcal{J},\bullet)}(\xi) \\ D(\xi) & Q(\xi) \end{bmatrix}$$

is not left unimodular. Thus, there exists a non-zero signal $\begin{bmatrix} l^* \\ u^* \end{bmatrix}$ that satisfies

$$0 = \begin{bmatrix} M^{(\mathcal{J},\bullet)}(\sigma) & P^{(\mathcal{J},\bullet)}(\sigma) \\ D(\sigma) & Q(\sigma) \end{bmatrix} \begin{bmatrix} l^* \\ u^* \end{bmatrix}.$$

Now, consider

$$\begin{bmatrix} y \\ 0 \end{bmatrix} = \begin{bmatrix} M(\sigma) & P(\sigma) \\ D(\sigma) & Q(\sigma) \end{bmatrix} \begin{bmatrix} l^* \\ u^* \end{bmatrix},$$

clearly $\|y\| \leq m + 1 - k$, this implies $\delta_c(\Sigma) \leq m + 1 - k$.

To prove $\delta_c(\Sigma) \geq m + 1 - k$, let $\tilde{y}$ be a signal such that $\begin{bmatrix} \tilde{y} \\ \tilde{u} \end{bmatrix} \in \mathcal{B}$ with $\|\tilde{y}\| = \delta_c(\Sigma)$. Thus, there exists a non-zero signal $\begin{bmatrix} \tilde{l} \\ \tilde{u}, \end{bmatrix}$ such that

$$\begin{bmatrix} \tilde{y} \\ 0 \end{bmatrix} = \begin{bmatrix} M(\sigma) & P(\sigma) \\ D(\sigma) & Q(\sigma) \end{bmatrix} \begin{bmatrix} \tilde{l} \\ \tilde{u} \end{bmatrix}.$$

Define $\bar{\mathcal{J}} \subseteq \{1, ..., m\}$ as the set of cardinality $\delta_c(\Sigma)$ for which $\tilde{y}^{(\mathcal{J})} = 0$, then

$$0 = \begin{bmatrix} M^{(\mathcal{J},\bullet)}(\sigma) & P^{(\mathcal{J},\bullet)}(\sigma) \\ D(\sigma) & Q(\sigma) \end{bmatrix} \begin{bmatrix} \tilde{l} \\ \tilde{u} \end{bmatrix}$$

must hold. Since $\begin{bmatrix} \tilde{l} \\ \tilde{u} \end{bmatrix} \neq 0$, then $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) & P^{(\mathcal{J},\bullet)}(\xi) \\ D(\xi) & Q(\xi) \end{bmatrix}$ is not left unimodular. This implies $k \geq m + 1 - \delta_s(\Sigma)$. The proof is now complete.

□

We are now ready to introduce the attack correction algorithm (see the following insert Algorithm 9) .

The following theorem provides a sufficient condition for a successful attack correction for a system under actuator and sensor attacks.

Algorithm 9 Attack correction for system under actuator and sensor attacks

1: **procedure** $(M(\xi), D(\xi), P(\xi), Q(\xi), r_s, u, \delta_c(\Sigma), \hat{\eta})$

▷ Given $M(\xi), D(\xi), P(\xi), Q(\xi), r_s, u, \delta_c(\Sigma)$, compute $\hat{\eta} = \begin{bmatrix} \hat{\eta}_s \\ \hat{\eta}_a \end{bmatrix}$.

2: For any subset $\mathcal{J} \in \{1, ..., m\}$ of cardinality $m + 1 - \delta_c(\Sigma)$, calculates the inverse matrix $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) & P^{(\mathcal{J},\bullet)}(\xi) \\ D(\xi) & Q(\xi) \end{bmatrix}^{-1}$.

3: Calculate

$$\begin{bmatrix} \hat{l} \\ \hat{u} \end{bmatrix} = \text{Maj}\left\{ \begin{bmatrix} M^{(\mathcal{J},\bullet)}(\sigma) & P^{(\mathcal{J},\bullet)}(\sigma) \\ D(\sigma) & Q(\sigma) \end{bmatrix}^{-1} \begin{bmatrix} r_s^{(\mathcal{J})} \\ 0 \end{bmatrix} \right\},$$

where the majority votes are taken over all subsets $\mathcal{J}$ of cardinality $m + 1 - \delta_c(\Sigma)$.

4: Calculate

$$\hat{\eta}_a = u - \hat{u}.$$

5: Calculate

$$\hat{\eta}_s = r_s - \begin{bmatrix} M(\sigma) & P(\sigma) \end{bmatrix} \begin{bmatrix} \hat{l} \\ \hat{u} \end{bmatrix}.$$

6: **return** $\hat{\eta} = \begin{bmatrix} \hat{\eta}_s \\ \hat{\eta}_a \end{bmatrix}$.

7: **end procedure**

Theorem 5.5. *Assume a strongly observable system Σ under actuator and sensor attacks as in* (5.3)*; then, Algorithm 9 is guaranteed to produce the correct attack signal, i.e., $\hat{\eta} = \eta$ if the sensor attack satisfies $\|\eta_s\| < \delta_c(\Sigma)/2$, and any actuator attacks.*

Proof. The proof of this theorem is divided into two parts:

a) We first prove that when the set $\mathcal{J}$ contains only attack-free sensors, then, the resulting votes regarding such chosen $\mathcal{J}$ satisfies $\hat{\eta} = \eta$. Denote the total number of such correct votes by T.

b) We then prove that the total number of votes for an incorrect signal must be smaller than T.

For **a)**, we first note that given $\|\eta_s\| < \delta_c(\Sigma)/2$, then $\delta_c(\Sigma) - 1 \geq \|\eta_s\|$ must hold. Let $\mathcal{J}$ be a subset of cardinality $p = m + 1 - \delta_s(\Sigma)$ from the set of attack-free sensors, then,

$$\begin{bmatrix} r_s^{(\mathcal{J})} \\ 0 \end{bmatrix} = \begin{bmatrix} y^{(\mathcal{J})} \\ 0 \end{bmatrix} = \begin{bmatrix} M^{(\mathcal{J},\bullet)}(\sigma) & P^{(\mathcal{J},\bullet)}(\sigma) \\ D(\sigma) & Q(\sigma) \end{bmatrix} \begin{bmatrix} l' \\ u - \eta_a \end{bmatrix}. \tag{5.7}$$

Given $\|\mathcal{J}\|_c = p$, then the observer $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) & P^{(\mathcal{J},\bullet)}(\xi) \\ D(\xi) & Q(\xi) \end{bmatrix}^{-1}$ is well-defined and we have

$$\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\sigma) & P^{(\mathcal{J},\bullet)}(\sigma) \\ D(\sigma) & Q(\sigma) \end{bmatrix}^{-1} \begin{bmatrix} r_s^{(\mathcal{J})} \\ 0 \end{bmatrix} = \begin{bmatrix} l' \\ u - \eta_a \end{bmatrix}. \tag{5.8}$$

Following step 4 and 5 in Algorithm 9, we can now trivially conclude that for such chosen $\mathcal{J}$, $\hat{\eta} = \eta$ must hold. There are a total of $\binom{m-\|\eta_s\|}{p}$ ways to choose a subset $\mathcal{J}$ of cardinality p from the set of attack-free sensors. Each choice leading to $\hat{l}$ as the correct result, thus we have $T = \binom{m-\|\eta_s\|}{p}$. This completes part **a)**.

For part **b)**, we need to prove that the number of votes for an incorrect signal must be a number smaller than $T = \binom{m-\|\eta_s\|}{p}$. We prove this by contradiction.

Let us assume that there exists another incorrect vote, say $\begin{bmatrix} l^* \\ u^* \end{bmatrix}$ that has at least T votes, then all involving indices $\mathcal{J}$ forms a set, say $\mathcal{F}^*$. Knowing that if we choose any subset $\mathcal{J} \subset \mathcal{F}^*$, with $\|\mathcal{J}\|_c = p$, the algorithm yields a unique vote. Put differently, any p indices from $\mathcal{F}^*$ leads to the same incorrect signal. In order for the incorrect signal to have at least T votes, the cardinality of F^* should be at least $m - \|\eta_s\|$.

Given that $\|\eta_s\| < \delta_c(\Sigma)/2$ and thus $2\|\eta_s\| < \delta_c(\Sigma)$, we have $m - \|\eta_s\| > m - \delta_c(\Sigma) + \|\eta_s\|$. This implies that there are at least $p = m - \delta_c(\Sigma) + 1$ attack-free sensors in $\mathcal{F}^\star$. Since we know that any p attack-free sensors leads to a correct vote $\begin{bmatrix} l' \\ u - \eta_a \end{bmatrix}$, then $\begin{bmatrix} l^* \\ u^* \end{bmatrix} = \begin{bmatrix} l' \\ u - \eta_a \end{bmatrix}$ must hold, which contradicts our assumption that $\begin{bmatrix} l^* \\ u^* \end{bmatrix} \neq \begin{bmatrix} l' \\ u - \eta_a \end{bmatrix}$. Thus, we conclude that there exists no such signal $\begin{bmatrix} l^* \\ u^* \end{bmatrix}$ that has at least T votes. This completes part **b)** of the proof. □

From the above correction Algorithm 9, we see that the left unimodularity of matrix $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) & P^{(\mathcal{J},\bullet)}(\xi) \\ D(\xi) & Q(\xi) \end{bmatrix}$ plays a crucial role in the correction algorithm.

Let us now recall the concept of strong observability as in [12,57] (or equivalently trivially secure for actuator attack as in Theorem 4.2), we see that if the polynomial matrix

$\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) & P^{(\mathcal{J},\bullet)}(\xi) \\ D(\xi) & Q(\xi) \end{bmatrix}$ is left unimodular, then the overall system $\begin{bmatrix} M(\xi) & P(\xi) \\ D(\xi) & Q(\xi) \end{bmatrix}$ must be strongly observable. This implies that the left unimodularity of $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) & P^{(\mathcal{J},\bullet)}(\xi) \\ D(\xi) & Q(\xi) \end{bmatrix}$ is an even stronger property as compare to the strong observability. This shows similarities when compared with the notion of $(s,r)-$sparse strong observability of [57], where s is upper bounded by q and r is upper bounded by m. More specifically, if we let $s = 0$, and $r = \lfloor \delta_c(\Sigma)/2 \rfloor$, then the left unimodularity of $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) & P^{(\mathcal{J},\bullet)}(\xi) \\ D(\xi) & Q(\xi) \end{bmatrix}$ is equivalent to the $(0,r)-$sparse strong observability.

The essences of left unimodularity for matrix $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) & P^{(\mathcal{J},\bullet)}(\xi) \\ D(\xi) & Q(\xi) \end{bmatrix}$ and the $(0,r)$ - sparse strong observability are the same. More specifically, it requires the system to maintain strongly observable even if certain available sensor output signals are removed from the system.

However, the key difference between our thesis and [57] lies inside the objective of attack correction. The work [57] does not attempt to achieve attack correction when $\eta_a \geq q/2$. However, in our case, our objective in this Section is to reconstruct η when all actuators are attacked together with some (but not all) sensors. This fundamental difference leads to a different approach and in effect, a different result.

In Section F of [22], a simulation result was presented for a 14-bus power network. The simulation result coincides with our correctability statement. From Figure (b) of [22], we see that the success rate of the attack correction is dominated by the number of sensor attacks, while the number of attacked actuators has much less effect on the success rate.

Comparing Algorithm 9 with the previously proposed sensor only attack correction Algorithm 5 and actuator only attack correction Algorithm 7, it can be seen that Algorithm 9 is based on strong observability, while the other algorithms do not require such property. This is attributed to the lack of prior attack knowledge. Since either sensor or actuator can be attacked in this case, the strong observability is vital for Algorithm 9.

5.2 Prior Sensor Knowledge

Let us now consider a situation wherein a system is under both actuator and sensor attacks, and we know that a subset of sensors is not accessible by the attacker. We denote such an assumption as a specific prior sensor knowledge. In this section, we assume that the system Σ is strongly observable, as seen in previous section 5.1. Based on such prior sensor knowledge, sensor and actuator attacks detection and correction methods are presented. To the best our knowledge, these results are additions to extant literature.

5.2.1 Preliminaries and Detectability

The attack-free system model we consider in this subsection is an ILO image representation, which is the same as the one mentioned in equation (3.15). The behaviour of the attack-free system is defined as in (3.18). In this subsection, prior sensor knowledge is considered where we assume the subset of sensors with index $\mathcal{I}_s \subsetneq \{1, ..., m\}$ is attack-free. It is noteworthy that if $\mathcal{I}_s = \{1, ..., m\}$, then we are in a different scenario, namely, actuator only attack, as seen in Chapter 4, where the assumption on strong observability is not necessarily needed. In other words, it is no longer suffices to treat the actuator only attack as a special case of sensor and actuator attacks with prior sensor knowledge. Knowing for sure there exist no sensor attack ($\mathcal{I}_s = \{1, ..., m\}$) or only processing certain prior sensor knowledge ($\mathcal{I}_s \subsetneq \{1, ..., m\}$) will lead to different approaches.

In our scope, a system under both sensor and actuator attacks with prior sensor knowledge $\mathcal{I}_s$ in the ILO image representation is given as follows:

$$\begin{bmatrix} r_s - \eta_s \\ 0 \end{bmatrix} = \begin{bmatrix} M(\sigma) & P(\sigma) \\ D(\sigma) & Q(\sigma) \end{bmatrix} \begin{bmatrix} l' \\ u - \eta_a \end{bmatrix}, \text{ where } \eta_s \text{ satisfies } \eta_s^{(\mathcal{I}_s)} = 0. \tag{5.9}$$

The behaviour of interest in this subsection is a subspace of $\mathcal{B}$, denoted as $\mathcal{B}_{ps}$ where

the subscript 'ps' stands for prior sensor knowledge, and is defined as follows:

$$\mathcal{B}_{ps} := \left\{ \left[\begin{array}{c} 0^{(\mathcal{I}_s)} \\ y^{(\bar{\mathcal{I}}_s)} \\ \hline u \end{array}\right] \mathbb{Z}_+ \to \mathbb{R}^{m+q} \,\middle|\, \left[\begin{array}{c} 0^{(\mathcal{I}_s)} \\ y^{(\bar{\mathcal{I}}_s)} \\ \hline u \end{array}\right] \in \mathcal{B} \right\}. \tag{5.10}$$

The attack detectability of an attack signal can be defined as follows:

A non-zero attack signal $\eta = \left[\begin{array}{c} 0^{(\mathcal{I}_s)} \\ \eta_s^{(\bar{\mathcal{I}}_s)} \\ \hline \eta_a \end{array}\right]$ is detectable if $\left[\begin{array}{c} 0^{(\mathcal{I}_s)} \\ \eta_s^{(\bar{\mathcal{I}}_s)} \\ \hline \eta_a \end{array}\right] \notin \mathcal{B}_{ps}$.

Now we are ready to propose the definition of a security index $\delta_{ps}(\Sigma)$ in this specific setup.

Definition 5.4. *Given the sensor set $\mathcal{I}_s$ is not accessible by the attacker, the sensor and actuator security index of the system Σ is defined as:*

$$\delta_{ps}(\Sigma) := \min_{0 \neq \left[\begin{array}{c} 0^{(\mathcal{I}_s)} \\ y^{(\bar{\mathcal{I}}_s)} \\ u \end{array}\right] \in \mathcal{B}_{ps}} \left\| \left[\begin{array}{c} y^{(\bar{\mathcal{I}}_s)} \\ u \end{array}\right] \right\|, \tag{5.11}$$

further define $\delta_{ps}(\Sigma) = \|\bar{\mathcal{I}}_s\|_c + 1$ if $\mathcal{B}_{ps} = \{0\}$.

Definition 5.5. *We call a system Σ trivially secure if $\delta_{ps} = \|\bar{\mathcal{I}}_s\|_c + 1$.*

To calculate the security index $\delta_{ps}(\Sigma)$, we first recall the kernel representation $R(\xi) \begin{bmatrix} y \\ u \end{bmatrix} = 0$. We then propose the following theorem.

Theorem 5.6. *The sensor and actuator security index with prior sensor knowledge $\mathcal{I}_s$ satisfies $\delta_{ps}(\Sigma) = k + 1$ where k is the largest integer in $\{1, ..., \|\bar{\mathcal{I}}_s\|_c\}$ such that for any submatrix $R^{(\bullet, \mathcal{J})}(\xi)$ with $\mathcal{J} \subseteq \{1, ..., m+q\}/\mathcal{I}_s$ and $\|\mathcal{J}\|_c = k$ is left unimodular.*

Proof. The proof follows as shown in Theorem 3.1. □

A sufficient condition for detectable attacks can then be stated as follows:

Theorem 5.7. *If the number of actuator and sensor attacks* $\left\| \begin{bmatrix} 0^{(\mathcal{I}_s)} \\ \eta_s^{(\bar{\mathcal{I}}_s)} \\ \hline \eta_a \end{bmatrix} \right\| < \delta_{ps}(\Sigma)$*, then such attack signal is detectable.*

Similar to the previous instances, Algorithm 3 and 4 can be used as the attack detection algorithms for a system under actuator and sensor attacks with prior sensor knowledge, replacing η_s by η. Moreover, we propose the following theorem:

Theorem 5.8. *Detection Algorithm 3 and 4 are guaranteed to achieve correct actuator and sensor attack detection with prior sensor knowledge* $\mathcal{I}_s$*, if the attack signal* η *is detectable or zero. A sufficient condition for a successful attack detection is:* $\|\eta\| < \delta_{ps}(\Sigma)$.

5.2.2 Attack Correction

Analogous to Definition 5.3, we now propose the definition for the correction index $\delta_{c-ps}(\Sigma)$ where the subscript 'c-ps' stands for correction index, prior sensor knowledge.

Definition 5.6. *The correction index* $\delta_{c-ps}(\Sigma)$ *of a system* Σ *with prior sensor knowledge* $\mathcal{I}_s$ *is defined as:*

$$\delta_{c-ps}(\Sigma) := \min_{0 \neq \left[\begin{smallmatrix} 0^{(\mathcal{I}_s)} \\ y^{(\bar{\mathcal{I}}_s)} \\ \hline u \end{smallmatrix}\right] \in \mathcal{B}_{ps}} \|y^{(\bar{\mathcal{I}}_s)}\|. \tag{5.12}$$

Further define $\delta_{c-ps}(\Sigma) = \|\bar{\mathcal{I}}_s\|_c + 1$ *if the system is trivially secure.*

In order to calculate the correction index $\delta_{c-ps}(\Sigma)$, we propose the following theorem.

Theorem 5.9. *Consider a system* Σ *in an ILO image representation as in* (5.9)*, then the correction index* $\delta_{c-ps}(\Sigma) = \|\bar{\mathcal{I}}_s\|_c + 1 - k$ *where k is the smallest integer in* $\{0, 1, ..., \|\bar{\mathcal{I}}_s\|_c\}$*such that for any subset* $\mathcal{J} \subseteq \bar{\mathcal{I}}_s$ *of cardinality k, the polynomial matrix* $\begin{bmatrix} M^{(\mathcal{I}_s,\bullet)}(\xi) & P^{(\mathcal{I}_s,\bullet)}(\xi) \\ M^{(\mathcal{J},\bullet)}(\xi) & P^{(\mathcal{J},\bullet)}(\xi) \\ D(\xi) & Q(\xi) \end{bmatrix}$ *is left unimodular.*

Proof. The proof follows as seen in Theorem 5.4. □

Theorem 5.10. *The system is trivially secure if $k = 0$, or equivalently, the polynomial matrix* $\begin{bmatrix} M^{(\mathcal{I}_s,\bullet)}(\xi) & P^{(\mathcal{I}_s,\bullet)}(\xi) \\ D(\xi) & Q(\xi) \end{bmatrix}$ *is left unimodular.*

We now present Algorithm 10 regarding the attack correction.

Algorithm 10 Actuator and sensor attacks correction for general system Σ with sensor prior knowledge

1: **procedure** $(M(\xi), D(\xi), P(\xi), Q(\xi), r_s, u, \mathcal{I}_s, \delta_{c-ps}(\Sigma), \hat{\eta})$

▷ Given $M(\xi), D(\xi), P(\xi), Q(\xi), r_s, u, \mathcal{I}_s, \delta_{c-ps}(\Sigma)$, compute $\hat{\eta} = \left[\begin{array}{c} 0^{(\mathcal{I}_s)} \\ \hline \hat{\eta}_s^{(\bar{\mathcal{I}}_s)} \\ \hline \hat{\eta}_a \end{array}\right]$.

2: For each subset $\mathcal{J} \subseteq \bar{\mathcal{I}}_s$ of cardinality $\|\bar{\mathcal{I}}_s\|_c + 1 - \delta_{c-ps}(\Sigma)$, calculates the inverse matrix $\begin{bmatrix} M^{(\mathcal{I}_s,\bullet)}(\xi) & P^{(\mathcal{I}_s,\bullet)}(\xi) \\ M^{(\mathcal{J},\bullet)}(\xi) & P^{(\mathcal{J},\bullet)}(\xi) \\ D(\xi) & Q(\xi) \end{bmatrix}^{-1}$.

3: Compute

$$\begin{bmatrix} \hat{l} \\ \hat{u} \end{bmatrix} = \text{Maj}\left\{ \begin{bmatrix} M^{(\mathcal{I}_s,\bullet)}(\sigma) & P^{(\mathcal{I}_s,\bullet)}(\sigma) \\ M^{(\mathcal{J},\bullet)}(\sigma) & P^{(\mathcal{J},\bullet)}(\sigma) \\ D(\sigma) & Q(\sigma) \end{bmatrix}^{-1} \left[\begin{array}{c} r_s^{(\mathcal{I}_s)} \\ \hline r_s^{(\mathcal{J})} \\ \hline 0 \end{array}\right] \right\},$$

where the majority vote is taken over all subsets $\mathcal{J}$ of cardinality $\|\bar{\mathcal{I}}_s\|_c + 1 - \delta_{c-ps}(\Sigma)$.

4: Compute

$$\hat{\eta}_a = u - \hat{u}.$$

5: Compute

$$\hat{\eta}_s = r_s - \begin{bmatrix} M(\sigma) & P(\sigma) \end{bmatrix} \begin{bmatrix} \hat{l} \\ \hat{u} \end{bmatrix}. \tag{5.13}$$

6: **return** $\hat{\eta} = \begin{bmatrix} \hat{\eta}_s \\ \hat{\eta}_a \end{bmatrix}$.

7: **end procedure**

The following theorem provides a sufficient condition for a successful attack correction concerning a system under actuator and sensor attacks with prior sensor knowledge.

Theorem 5.11. *Assume a strongly observable system Σ under actuator and sensor attacks with prior sensor knowledge $\mathcal{I}_s$ as in* (5.9)*, Algorithm 10 is guaranteed to produce the correct attack signal, i.e., $\hat{\eta} = \eta$ if the sensor attack satisfies $\|\eta_s\| < \delta_{c-ps}(\Sigma)/2$ and any actuator attacks.*

Furthermore, if $\delta_{c-ps} = \|\bar{\mathcal{I}}_s\|_c + 1$, *then any* η *can be corrected.*

Proof. The proof follows as mentioned in Theorem 5.5 □

5.3 Prior Actuator Knowledge

Let us now consider the situation when a system is under both actuator and sensor attacks, and we are given a subset of actuators that is not accessible by the attacker. We denote such an assumption as a specific prior actuator knowledge. Attack detection and correction methods are presented. In this section, the objective of attack correction is the same as mentioned in the previous sections. We are aiming to achieve guaranteed attack correction when all the accessible actuators and some sensors are under attack. Based on such correction objective, it is not necessary to assume that the entire system is strongly observable. However, if we are looking at the subsystem without those inaccessible actuators, then strong observability is vital for that subsystem. It is for this reason that we assume that the subsystem of Σ by removing the inaccessible actuators is strongly observable. To the best of our knowledge, the results in this section are new.

5.3.1 Preliminaries and Detectability

The attack-free system model we consider in this subsection is an ILO image representation, which is the same as seen in equation (3.15). Meanwhile behaviour of the attack-free system is defined as mentioned in (3.18). In this subsection, prior actuator knowledge is considered where we assume that the subset of actuators with index $\mathcal{I}_a \subsetneq \{1, ..., q\}$ is attack-free. Similar as in Section 5.2, if $\mathcal{I}_a = \{1, ..., q\}$, we are then in a different scenario, namely, sensor only attack as in Chapter 3. In Chapter 3, strong observability of a system is not necessarily needed to perform attack detection and correction, which means that is no longer suffices to treat the sensor only attack as a special case of sensor and actuator attacks case.

In our scope, a system under both sensor and actuator attacks with prior actuator

knowledge $\mathcal{I}_a$ is given as follows:

$$\begin{bmatrix} r_s - \eta_s \\ 0 \end{bmatrix} = \begin{bmatrix} M(\sigma) & P(\sigma) \\ D(\sigma) & Q(\sigma) \end{bmatrix} \begin{bmatrix} l' \\ u - \eta_a \end{bmatrix}, \text{ where } \eta_a \text{ satisfies } \eta_a^{(\mathcal{I}_a)} = 0. \tag{5.14}$$

The behaviour of interest in this subsection is a subspace of $\mathcal{B}$, denoted as $\mathcal{B}_{pa}$, defined as follows:

$$\mathcal{B}_{pa} := \left\{ \left[\begin{array}{c} y \\ \hline 0^{(\mathcal{I}_a)} \\ u^{(\bar{\mathcal{I}}_a)} \end{array}\right] : \mathbb{Z}_+ \to \mathbb{R}^{m+q} \,\middle|\, \left[\begin{array}{c} y \\ \hline 0^{(\mathcal{I}_a)} \\ u^{(\bar{\mathcal{I}}_a)} \end{array}\right] \in \mathcal{B} \right\}. \tag{5.15}$$

The attack detectability of an attack signal is defined as follows:

A non-zero attack signal $\eta = \left[\begin{array}{c} \eta_s \\ \hline 0^{(\mathcal{I}_a)} \\ \eta_a^{(\bar{\mathcal{I}}_a)} \end{array}\right]$ is detectable if $\left[\begin{array}{c} \eta_s \\ \hline 0^{(\mathcal{I}_a)} \\ \eta_a^{(\bar{\mathcal{I}}_a)} \end{array}\right] \notin \mathcal{B}_{pa}$.

We are now ready to propose the definition of a security index $\delta_{pa}(\Sigma)$ in this specific setup.

Definition 5.7. *Given in prior that the actuator set $\mathcal{I}_a$ is not accessible by the attacker, the sensor and actuator security index of the system Σ is defined as:*

$$\delta_{pa}(\Sigma) := \min_{0 \neq \left[\begin{array}{c} y \\ \hline 0^{(\mathcal{I}_a)} \\ u^{(\bar{\mathcal{I}}_a)} \end{array}\right] \in \mathcal{B}_{pa}} \left\| \begin{bmatrix} y \\ u^{(\bar{\mathcal{I}}_a)} \end{bmatrix} \right\|. \tag{5.16}$$

The trivially secure case is not applicable in this setup, which means that $\mathcal{B}_{pa} \neq \{0\}$ must hold, since the input signal $u^{(\bar{\mathcal{I}}_a)}$ can be arbitrary.

To calculate the security index $\delta_{pa}(\Sigma)$, we first recall the kernel representation $R(\xi) \begin{bmatrix} y \\ u \end{bmatrix} = 0$, and then propose the following theorem.

Theorem 5.12. *The sensor and actuator security index with prior actuator knowledge $\mathcal{I}_a$ satisfies $\delta_{pa}(\Sigma) = k + 1$ where k is the largest integer in $\{1, ..., m\}$ such that any submatrix $R^{(\bullet, \mathcal{J})}(\xi)$ with $\mathcal{J} \subseteq \{1, ..., m+q\} / \mathcal{I}_a$ and $\|\mathcal{J}\|_c = k$ is left unimodular.*

Proof. The proof follows as mentioned in Theorem 3.1. □

A sufficient condition for detectable sensor and actuator attacks with prior actuator knowledge $\mathcal{I}_a$ can then be stated as follows:

Theorem 5.13. *If the number of actuator and sensor attacks* $\left\| \begin{bmatrix} \eta_s \\ \hline 0^{(\mathcal{I}_a)} \\ \eta_a^{(\bar{\mathcal{I}}_a)} \end{bmatrix} \right\| < \delta_{pa}(\Sigma)$*, then such attack signal is detectable.*

Algorithm 3 and 4 can be used as the attack detection algorithms for a system under actuator and sensor attacks with prior actuator knowledge, replacing η_s by η. Moreover, we propose the following theorem:

Theorem 5.14. *Detection Algorithm 3 and 4 are guaranteed to achieve correct actuator and sensor attacks detection with prior actuator knowledge $\mathcal{I}_a$ if the attack signal η is detectable or zero. A sufficient condition for a successful attack detection is: $\|\eta\| < \delta_{pa}(\Sigma)$.*

5.3.2 Attack Correction

Analogous to Definition 5.3 and Definition 5.6, we now propose the definition for the correction index $\delta_{c-pa}(\Sigma)$ where the subscript 'c-pa' denotes the correction index, prior actuator knowledge.

$$\delta_{c-pa}(\Sigma) := \min_{0 \neq \begin{bmatrix} y \\ \hline 0^{(\mathcal{I}_a)} \\ u^{(\bar{\mathcal{I}}_a)} \end{bmatrix} \in \mathcal{B}_{pa}} \|y\| . \tag{5.17}$$

To calculate the correction index $\delta_{c-pa}(\Sigma)$, we propose the following theorem

Theorem 5.15. *Consider a system Σ in an ILO image representation as in equation* (5.14)*. The correction index then equals $\delta_{c-pa}(\Sigma) = m + 1 - k$, where k is the smallest integer in $\{0, 1, ..., m\}$ such that for any subset $\mathcal{J} \subseteq \{1, ..., m\}$ of cardinality k, the $(k + n) \times (q + \|\bar{\mathcal{I}}_a\|)$ polynomial*

matrix

$$\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) & P^{(\mathcal{J},\bar{\mathcal{I}}_a)}(\xi) \\ D(\xi) & Q^{(\bullet,\bar{\mathcal{I}}_a)}(\xi) \end{bmatrix}$$

is left unimodular.

Proof. Clearly, there exists a subset $\mathcal{J}$ of cardinality $k-1$ such that the polynomial matrix $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) & P^{(\mathcal{J},\bar{\mathcal{I}}_a)}(\xi) \\ D(\xi) & Q^{(\bullet,\bar{\mathcal{I}}_a)}(\xi) \end{bmatrix}$ is not left unimodular, which means that there exists a non-zero signal $\begin{bmatrix} \tilde{l} \\ \tilde{u}^{(\bar{\mathcal{I}}_a)} \end{bmatrix}$ with $\tilde{u}^{(\bar{\mathcal{I}}_a)} : \mathbb{Z}_+ \to \mathbb{R}^{\|\bar{\mathcal{I}}_a\|}$ that satisfies

$$0 = \begin{bmatrix} M^{(\mathcal{J},\bullet)}(\sigma) & P^{(\mathcal{J},\bar{\mathcal{I}}_a)}(\sigma) \\ D(\sigma) & Q^{(\bullet,\bar{\mathcal{I}}_a)}(\sigma) \end{bmatrix} \begin{bmatrix} \tilde{l} \\ \tilde{u}^{(\bar{\mathcal{I}}_a)} \end{bmatrix}.$$

Now define $\tilde{u} = \begin{bmatrix} \tilde{u}^{(\mathcal{I}_a)} \\ \tilde{u}^{(\bar{\mathcal{I}}_a)} \end{bmatrix}$ where $\tilde{u}^{(\mathcal{I}_a)} = 0$ and consider $\begin{bmatrix} y \\ 0 \end{bmatrix} = \begin{bmatrix} M(\sigma) & P(\sigma) \\ D(\sigma) & Q(\sigma) \end{bmatrix} \begin{bmatrix} \tilde{l} \\ \tilde{u} \end{bmatrix}$. Clearly, $\|y\| \leq m+1-k$ and this implies $\delta_{c-pa}(\Sigma) \leq m+1-k$.

To prove $\delta_{c-pa}(\Sigma) \geq m+1-k$, let $\hat{y}$ be a signal such that $\begin{bmatrix} \hat{y} \\ \hat{u} \end{bmatrix} \in \mathcal{B}$ with $\hat{u}^{(\mathcal{I}_a)} = 0$ and $\|\hat{y}\| = \delta_{c-pa}(\Sigma)$. Such signal $\begin{bmatrix} \hat{y} \\ \hat{u} \end{bmatrix} \in \mathcal{B}_{pa}$ exists due to equation (5.17); then, there exists a non-zero signal $\begin{bmatrix} \hat{l} \\ \hat{u} \end{bmatrix}$ with $\hat{u}^{(\mathcal{I}_a)} = 0$ satisfies

$$\begin{bmatrix} \hat{y} \\ 0 \end{bmatrix} = \begin{bmatrix} M(\sigma) & P(\sigma) \\ D(\sigma) & Q(\sigma) \end{bmatrix} \begin{bmatrix} \hat{l} \\ \hat{u} \end{bmatrix}.$$

Define $\bar{\mathcal{J}}$ as the set of cardinality $\delta_{c-pa}(\Sigma)$ for which $\hat{y}^{(\mathcal{J})} = 0$, then

$$0 = \begin{bmatrix} M^{(\mathcal{J},\bullet)}(\sigma) & P^{(\mathcal{J},\bar{\mathcal{I}}_a)}(\sigma) \\ D(\sigma) & Q^{(\bullet,\bar{\mathcal{I}}_a)}(\sigma) \end{bmatrix} \begin{bmatrix} \tilde{l} \\ \tilde{u}^{(\bar{\mathcal{I}}_a)} \end{bmatrix}.$$

Since $\begin{bmatrix} \hat{l} \\ \hat{u}^{(\mathcal{I}_a)} \end{bmatrix} \neq 0$, then $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) & P^{(\mathcal{J},\bar{\mathcal{I}}_a)}(\xi) \\ D(\xi) & Q^{(\bullet,\bar{\mathcal{I}}_a)}(\xi) \end{bmatrix}$ is not left unimodular. This implies $k \geq m + 1 - \delta_{c-pa}(\Sigma)$. This, in turn, completes the proof. □

Each inverse matrix $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) & P^{(\mathcal{J},\bar{\mathcal{I}}_a)}(\xi) \\ D(\xi) & Q^{(\bullet,\bar{\mathcal{I}}_a)}(\xi) \end{bmatrix}^{-1}$ can be interpreted as an observer, and the existence of such observer is related to the left unimodularity of $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) & P^{(\mathcal{J},\bar{\mathcal{I}}_a)}(\xi) \\ D(\xi) & Q^{(\bullet,\bar{\mathcal{I}}_a)}(\xi) \end{bmatrix}$. It can be seen that sufficient conditions for $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) & P^{(\mathcal{J},\bar{\mathcal{I}}_a)}(\xi) \\ D(\xi) & Q^{(\bullet,\bar{\mathcal{I}}_a)}(\xi) \end{bmatrix}$ being left unimodular is the bigger matrix $\begin{bmatrix} M(\xi) & P^{(\bullet,\bar{\mathcal{I}}_a)}(\xi) \\ D(\xi) & Q^{(\bullet,\bar{\mathcal{I}}_a)}(\xi) \end{bmatrix}$ being left unimodular. Now, consider an attack-free sub-system by removing the influences for inputs with indices $\mathcal{I}_a$, after which we have

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} M(\sigma) & P^{(\bullet,\bar{\mathcal{I}}_a)}(\sigma) \\ D(\sigma) & Q^{(\bullet,\bar{\mathcal{I}}_a)}(\sigma) \end{bmatrix} \begin{bmatrix} l \\ u^{(\bar{\mathcal{I}}_a)} \end{bmatrix}, \tag{5.18}$$

where $\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} P^{(\bullet,\mathcal{I}_a)}(\sigma) \\ Q^{(\bullet,\mathcal{I}_a)}(\sigma) \end{bmatrix} u^{(\mathcal{I}_a)}$. We can now conclude that a necessary condition for the existence of observer $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) & P^{(\mathcal{J},\bar{\mathcal{I}}_a)}(\xi) \\ D(\xi) & Q^{(\bullet,\bar{\mathcal{I}}_a)}(\xi) \end{bmatrix}^{-1}$ is the strong observability of subsystem (5.18).

We are now ready to introduce the attack correction algorithm (see the following insert Algorithm 11) for system under actuator and sensor attacks with prior actuator knowledge.

The essence of the correction algorithm can be summarised as follows: Given that certain actuator sets are attack-free, we first subtract the output increment produced by these attack-free inputs, as shown in equation (5.19). After this step, we transfer the attack correction problem with prior actuator knowledge to a correction problem without prior knowledge as in Section 5.1.2. Step 3-6 in Algorithm 11 is identical to Algorithm 9, Step 3-6 .

The following theorem provides a sufficient condition for a successful attack correc-

Algorithm 11 Attack correction for system under actuator and sensor attack with actuator prior knowledge

1: **procedure** $(M(\xi), D(\xi), P(\xi), Q(\xi), r_s, u, \mathcal{I}_a, \delta_{c-pa}(\Sigma), \hat{\eta})$

▷ Given $M(\xi), D(\xi), P(\xi), Q(\xi), r_s, u, \mathcal{I}_a, \delta_{c-pa}(\Sigma)$, compute $\hat{\eta} = \begin{bmatrix} \hat{\eta}_s \\ \hline 0^{(\mathcal{I}_a)} \\ \hat{\eta}_a^{(\bar{\mathcal{I}}_a)} \end{bmatrix}$.

2: Calculate

$$s = \begin{bmatrix} s_a \\ s_b \end{bmatrix} = \begin{bmatrix} r_s \\ 0 \end{bmatrix} - \begin{bmatrix} P^{(\bullet, \mathcal{I}_a)}(\sigma) \\ Q^{(\bullet, \mathcal{I}_a)}(\sigma) \end{bmatrix} u^{(\mathcal{I}_a)}. \tag{5.19}$$

3: Calculate

$$\begin{bmatrix} \hat{l} \\ \hat{u}^{(\mathcal{I}_a)} \end{bmatrix} = \text{Maj} \left\{ \begin{bmatrix} M^{(\mathcal{J}, \bullet)}(\sigma) & P^{(\mathcal{J}, \bar{\mathcal{I}}_a)}(\sigma) \\ D(\sigma) & Q^{(\bullet, \bar{\mathcal{I}}_a)}(\sigma) \end{bmatrix}^{-1} \begin{bmatrix} s_a^{(\mathcal{J})} \\ s_b \end{bmatrix} \right\},$$

where the majority vote is taken over all subset $\mathcal{J} \subseteq \{1, ..., m\}$ of cardinality $\|\mathcal{J}\|_c = m + 1 - \delta_{c-pa}(\Sigma)$.

4: Calculate

$$\hat{\eta}_a = \begin{bmatrix} 0^{(\mathcal{I}_a)} \\ u^{(\bar{\mathcal{I}}_a)} - \hat{u}^{(\bar{\mathcal{I}}_a)} \end{bmatrix}.$$

5: Calculate

$$\hat{\eta}_s = r_s - \begin{bmatrix} M(\sigma) & P(\sigma) \end{bmatrix} \begin{bmatrix} \hat{l} \\ \hline u^{(\mathcal{I}_a)} \\ \hat{u}^{(\bar{\mathcal{I}}_a)} \end{bmatrix}. \tag{5.20}$$

6: **return** $\hat{\eta} = \begin{bmatrix} \hat{\eta}_s \\ \hat{\eta}_a \end{bmatrix}$.

7: **end procedure**

tion.

Theorem 5.16. *Assume a system Σ under actuator and sensor attacks with prior actuator knowledge as in* (5.14)*, further assume the sub-system of Σ by removing the influences for inputs with indices $\mathcal{I}_a$ as mentioned in* (5.18) *is strongly observable, then Algorithm 11 is guaranteed to produce the correct attack signal, i.e., $\hat{\eta} = \eta$ if the sensor attack satisfies $\|\eta_s\| < \delta_{c-pa}(\Sigma)/2$ and any actuator attacks.*

Proof. The proof follows as in Theorem 4.7.

□

Before we move on to the final topic in this chapter, it is worth mentioning that the left unimodularity of $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) & P^{(\mathcal{J},\bar{\mathcal{I}}_a)}(\xi) \\ D(\xi) & Q^{(\bullet,\bar{\mathcal{I}}_a)}(\xi) \end{bmatrix}$, or equivalently, the strong observability for the subsystem by neglecting actuators $\mathcal{I}_a$, is incompatible with the concept of (s,r) - sparse strong observability as in [57].

In fact, the left unimodularity of $\begin{bmatrix} M^{(\mathcal{J},\bullet)}(\xi) & P^{(\mathcal{J},\bar{\mathcal{I}}_a)}(\xi) \\ D(\xi) & Q^{(\bullet,\bar{\mathcal{I}}_a)}(\xi) \end{bmatrix}$ is a weaker condition. This is because it only requires the matrix regarding the specific index $\bar{\mathcal{I}}_a$ being left unimodular while the (s,r) - sparse strong observability requires all the submatrices of the same size are left unimodular. The weakness comes from the fact that in our setup, we know set $\mathcal{I}_a$ whereas [57] does not make that assumption.

5.4 Prior Actuator and Sensor Knowledge

In this section, we briefly discuss the problem of attack detection and attack correction for a system under both actuator and sensor attacks with prior sensor and actuator knowledge. The idea for solving detection and correction issue is as follows: we first take into account the prior actuator knowledge. Subtract the increment caused by these known and attack-free inputs at the output. Then, the remaining subsystem can be treated as a system under sensor and actuator attacks with only prior sensor knowledge, as in Section 5.2.

Consider the following attacked system model:

$$\begin{bmatrix} r_s - \eta_s \\ 0 \end{bmatrix} = \begin{bmatrix} M(\sigma) & P(\sigma) \\ D(\sigma) & Q(\sigma) \end{bmatrix} \begin{bmatrix} l' \\ u - \eta_a \end{bmatrix}, \text{ where } \begin{bmatrix} \eta_s \\ \eta_a \end{bmatrix} \text{ satisfies } \begin{bmatrix} \eta_s^{(\mathcal{I}_s)} \\ \eta_a^{(\mathcal{I}_a)} \end{bmatrix} = 0. \tag{5.21}$$

Using a similar idea as proposed in Section 5.3.2, we can see that given the attack-free actuator set $\mathcal{I}_a$, the attacked system model can be expressed as follows:

$$\begin{bmatrix} r_s \\ 0 \end{bmatrix} - \begin{bmatrix} P^{(\bullet,\mathcal{I}_a)}(\sigma) \\ Q^{(\bullet,\mathcal{I}_a)}(\sigma) \end{bmatrix} u^{(\mathcal{I}_a)} - \begin{bmatrix} \eta_s \\ 0 \end{bmatrix} = \begin{bmatrix} M(\sigma) & P^{(\bullet,\bar{\mathcal{I}}_a)}(\sigma) \\ D(\sigma) & Q^{(\bullet,\bar{\mathcal{I}}_a)}(\sigma) \end{bmatrix} \begin{bmatrix} l' \\ u^{(\bar{\mathcal{I}}_a)} - \eta_a^{(\bar{\mathcal{I}}_a)} \end{bmatrix}. \tag{5.22}$$

Since r_s, u and the system matrices are given, we compute signal $\begin{bmatrix} r_a \\ r_b \end{bmatrix}$ as follows:

$$\begin{bmatrix} r_a \\ r_b \end{bmatrix} = \begin{bmatrix} r_s \\ 0 \end{bmatrix} - \begin{bmatrix} P^{(\bullet,\mathcal{I}_a)}(\sigma) \\ Q^{(\bullet,\mathcal{I}_a)}(\sigma) \end{bmatrix} u^{(\mathcal{I}_a)},$$

then the attacked system (equation (5.22)) can be written as:

$$\begin{bmatrix} r_a - \eta_s \\ r_b \end{bmatrix} = \begin{bmatrix} M(\sigma) & P^{(\bullet,\bar{\mathcal{I}}_a)}(\sigma) \\ D(\sigma) & Q^{(\bullet,\bar{\mathcal{I}}_a)}(\sigma) \end{bmatrix} \begin{bmatrix} l' \\ u^{(\bar{\mathcal{I}}_a)} - \eta_a^{(\bar{\mathcal{I}}_a)} \end{bmatrix}. \tag{5.23}$$

Note that equation (5.23) can be recognised as a system under actuator and sensor attacks with only prior sensor knowledge. This type of problem is explained and solved in the previous section 5.2. Therefore, a detailed explanation is no longer necessary.

We use the following table to compare and summarise the contents of this chapter.

Table 5.1: Sufficient conditions for sensor and actuator attacks detection and correction

Without prior knowledge	With prior sensor knowledge	With prior actuator knowledge
Security index: $\delta_{sa}(\Sigma) := \min_{0 \neq \begin{bmatrix} y \\ u \end{bmatrix} \in \mathcal{B}} \left\| \begin{bmatrix} y \\ u \end{bmatrix} \right\|$	Security index: $\delta_{ps}(\Sigma) := \min_{0 \neq \begin{bmatrix} 0^{(\mathcal{I}_s)} \\ y^{(\bar{\mathcal{I}}_s)} \\ u \end{bmatrix} \in \mathcal{B}_{ps}} \left\| \begin{bmatrix} y^{(\mathcal{I}_s)} \\ u \end{bmatrix} \right\|$	Security index: $\delta_{pa}(\Sigma) := \min_{0 \neq \begin{bmatrix} y \\ 0^{(\mathcal{I}_a)} \\ u^{(\bar{\mathcal{I}}_a)} \end{bmatrix} \in \mathcal{B}_{pa}} \left\| \begin{bmatrix} y \\ 0^{(\mathcal{I}_a)} \\ u^{(\mathcal{I}_a)} \end{bmatrix} \right\|$
Correction index: $\delta_c(\Sigma) := \min_{0 \neq \begin{bmatrix} y \\ u \end{bmatrix} \in \mathcal{B}} \|y\|$	Correction index: $\delta_{c-ps}(\Sigma) := \min_{0 \neq \begin{bmatrix} 0^{(\mathcal{I}_s)} \\ y^{(\bar{\mathcal{I}}_s)} \\ u \end{bmatrix} \in \mathcal{B}_{ps}} \|y^{(\bar{\mathcal{I}}_s)}\|$	Correction index: $\delta_{c-pa}(\Sigma) := \min_{0 \neq \begin{bmatrix} y \\ 0^{(\mathcal{I}_a)} \\ u^{(\bar{\mathcal{I}}_a)} \end{bmatrix} \in \mathcal{B}_{pa}} \|y\|$
Trivially secure N/A	Non-trivially secure $\delta_{ps}(\Sigma) = \|\bar{\mathcal{I}}_s\|_c + 1$	Trivially secure N/A
Detectability N/A	Detectability $\|\eta\| < \delta_{ps}(\Sigma)$	Detectability N/A
Correctability N/A	Correctability $\|\eta_s\| < \delta_{c-ps}(\Sigma)$	Detectability N/A
Trivially secure N/A	Non-trivially secure $\delta_{ps}(\Sigma) < \|\bar{\mathcal{I}}_s\|_c + 1$	Non-trivially secure Detectability $\|\eta_s\| < \delta_{pa}(\Sigma)$
Detectability N/A	Detectability $\|\eta\| < \delta_{ps}(\Sigma)$	Detectability $\|\eta_s\| < \delta_{pa}(\Sigma)$
Correctability N/A	Correctability $\|\eta_s\| < \delta_{c-ps}(\Sigma)/2$	Correctability $\|\eta_s\| < \delta_{c-pa}(\Sigma)/2$

5.5 Recapitulation

In this chapter, we have studied LTI CPS security under sensor and actuator attacks. The main points discussed were:

- To address the problem of attack detection, we extended the concept of security index in this actuator and sensor attacks case. A sensor and actuator security index $\delta_{sa}(\Sigma)$ were proposed.
- The objective of attack correction in this chapter is to reconstruct the sensor and actuator attacks signals when all actuators are corrupted together with some (but not all) attacked sensors.
- To achieve the above correction objective, the notion of strong observability plays an important role in this section.
- Unlike previous chapters, a different representation-free system parameter 'correction index' was proposed to address attack correction.
- Attack detection and correction methods for systems with prior sensor and/or actuator knowledge were discussed.

Chapter 6

An Example of Sensor Attack Detection and Correction Under Measurement Noise

In this chapter, we illustrate the working of the proposed sensor attack detection and correction methods in Chapter 3 via a particular engineering example, namely, a speed measurement system for a self-driving farming vehicle. In this chapter, the measurement signals are assumed to be noisy.

6.1 Background

6.1.1 Development of the Self-driving Farming Vehicle

The idea of a self-driving farming vehicle has been around since as early as 1940 when Frank W. Andrew invented his own driverless tractor [1]. In the 1950s, Ford developed a driverless tractor known as 'The Sniffer'. The development of this self-driving farming vehicle took place much earlier as compared to the autonomous road vehicle which first appeared in the 1980s. There are several reasons for such an earlier approach. 1, the tractors travel at a slower speed compared with road vehicles; this allows more reaction time for the electrical control unit. 2, the traffic conditions are simpler on a farm, i.e., fewer traffic rules, no lines on private land. 3, tractors operate in a predetermined enclosed area. With the advancement of science and technology, the popularity of agricultural automation and self driving farming vehicle has been increasing at a very fast pace in recent years. The work [39] provides a brief review regarding the recent development of

self driving farming vehicles. In [3], the network structure for smart precision farming is discussed. Apart from the growing academic research, there are also several primary manufacturers that actively seeking to produce driverless farming vehicles. Those manufactures include John Deere, ATC (Autonomous Tractor Corp), Fendt and CaseIH etc. Many of the produces are widely used on some crop farms. Farmers' demand for automated farming vehicle is also increasing in recent years.

6.1.2 Speed Measurement System

The vehicle speed measurement system of a self-driving vehicle is indeed a CPS. Various sensors provide measurement redundancy for the speed signal. The electrical control unit then performs a certain control mechanism based on such measurement signals. In this chapter, we consider three types of measurements that are directly or indirectly measuring the speed of the travelling farming vehicle. Direct speed measurements considered in this chapter include the wheel speed sensors and the ground speed sensors. Indirect speed measurement taken into consideration includes the Local Positioning System (LPS).

Unlike previous chapters, measurement noise is being considered in this chapter. The terminology 'measurement noise' (see e.g [4], [48]) is often used in the control community for describing a measurement error due to the finite resolution of any sensor, or measurement device, as well as the limitations of the measurement transducer (non-linearity, hysteresis etc.).

For each sensor being considered, we model the measurement noise as an additive bounded noise. We further assume that the measurement noise does not introduce any bias to the measurement, which means that the sensors have been calibrated to the driving condition. In many engineering applications, measurement noise is only calibrated with a relative noise against full-scale measurements, i.e., $x\%$ of the maximum measurement range, and thus, it is reasonable to make the boundedness assumption. In [5, 23, 27, 54, 56], the authors also take such a bounded noise approach. Apart from the unbiased and boundedness assumptions, we make no further assumptions regarding the distribution or statistics of the noise.

6.1.3 Relevant Literature

The first result on state estimation with bounded noise dates back to 1968 [54]. In [54], the state estimation process returns a time-varying ellipsoid which is guaranteed to contain the actual state. Our work in this chapter shows similarities compared with [54], in the sense that when no attack occurred, the actual state of the system lies inside a certain measurement region. The difference is that we use a system's behavioural approach; thus, the concept of measurement region is a signal space rather than time-varying ellipsoids. In both our work and [54], there is no guarantee to single out the exact state signal; nevertheless, the measurement region or estimated ellipsoid provides a good insight for state estimation.

Another major difference between the work in this chapter and [54] is that in addition to the bounded measurement noise, we also consider sensor attack signals. In recent work [44], the authors consider the state estimation problem under measurement noise and attacks. The state estimation process in [44] is discussed based on an l_0-based state estimation problems $P_{0,w}$. The work [54] further simplifies such an NP hard problem using an l_1-based state estimation problems $P_{1,w}$, as in [22]. In this chapter, in order to address the problem of attack detection and correction, we extend the attack detection and observer-based attack correction algorithm in Chapter 3 for this particular noisy example.

The problem of state estimation or system parameter identification under bounded noise is sometimes referred to as a membership set estimation problem. The noise can be expressed within a certain bounded set. The estimated state can also be described within a closed set. Such problems are fundamental problems in the control community, see for example [23],[64], also in fault detection literature [5], [58], [32].

The works in [31, 32] presented two theorems which describe the evolution of the state uncertainty set under bounded measurement uncertainty for attack-free SISO LTI systems. The proposed theorems yield an efficient algorithm for recursively updating the uncertainty state sets. These works are targeted for general dynamical SISO systems, while the work in this chapter focuses on a much simpler speed measurement system but with attacks. In this chapter, we provide a conceptual discussion around CPS security for this particular speed measurement system of a self-driving farming vehicle. The content

in this chapter serves as a starting point for future research pertaining to general noisy systems using the previously proposed detection/correction methods.

6.1.4 Objective of Research

The objectives in this chapter are to explore the attack detection/correction capabilities of this specific speed measurement system, illustrating some of the results from the previous chapters and outlining new ideas for future methods that deal with measurement noise.

In order to clearly illustrate the previously proposed concept of security index, we first assume a certain prior sensor knowledge of the system, namely, we assume that the positioning sensor (LPS) is guaranteed to be attack-free and that the attacker can only access the speed sensors. Under such prior knowledge assumption, we aim to extend our previous developed detection/correction methods in Section 3.3 to this noisy system. We also look at the case when we do not have such prior sensor knowledge. As a result, the security index of the system will decrease significantly without such prior sensor knowledge, as will be seen later on in this chapter.

The objective of attack detection is to determine whether an attack signal has occurred; the objective of attack correction is to produce an estimated speed signal which is close to the attack-free signal. As seen in previous chapters, the detectability and correctability are stated in terms of the sensor security index (with or without the prior sensor knowledge).

Another interesting topic regarding this engineering example is that, in many situations, the speed of the vehicle is an important safety parameter that directly reflects whether the vehicle is travelling hazardously. On the other hand, the position signal often acts as an auxiliary signal when measuring the speed of the vehicle. In terms of the content in this chapter, we say that reconstructing the correct speed signal is more important than reconstructing the position signal from a safety aspect. At a later stage in this chapter, we will briefly discuss how to take into account such correction priority (or in other words, unequal attack correction) and then propose an attack correction method to reconstruct the speed signal only.

6.2 System Model

For the sake of simplicity, we consider the farming vehicle is moving at a constant speed with a constant direction. The maximum speed [2] of a full-load tractor under working conditions is 10km/hour. In this chapter, we assume the speed of the tractor is v, where v is a constant signal smaller than 2.7m/s (= 10km/hour). Various sensors are measuring the speed directly or indirectly (via local positioning system, for example). For simplicity purposes, we consider three types of speed measurement sensors that are typically installed on farming vehicles, namely:

- Wheel Speed Sensor $v_w : \mathbb{Z}_+ \to \mathbb{R}$;
- Ground Speed Sensor $v_g : \mathbb{Z}_+ \to \mathbb{R}$;

  Both v_w and v_g provide direct speed measurement. The unit of the measurements are m/s.

- Local positioning system (LPS) $p_x : \mathbb{Z}_+ \to \mathbb{R}$.

  We assume p_x is the vehicle's distance in meters to a certain reference point on the 1-dimensional line. The unit of p_x is m.

The received signals provided by those sensors are not ideal. In reality, sensor measurements are typically having different bias and different drift parameters. For example, a wheel slippage will cause a positive bias for the wheel speed sensor; the environment conditions (temperature, atmospheric pressure etc.) will lead to a different bias for the ground speed sensor. In this chapter, for the sake of simplicity, we assume the measurement signals are pre-calibrated and unbiased.

The measurement noise influences the output signals. As mentioned before, we model the measurement noise as an additive bounded noise. To illustrate, if we denote the actual speed of the farming vehicle by v, then the measurement signal for the wheel speed sensor satisfies

$$v_w = v + w_w, \tag{6.1}$$

where w_w denotes a bounded noise that satisfies $|w_w(t)| \leq \alpha_w(t)$ for all $t \in \mathbb{Z}_+$. Signal α_w is a constant signal, which can be interpreted as an **error margin** of the wheel speed mea-

surements. More specifically, if the wheel speed sensor provides a speed measurement v_w, then the actual speed v satisfies

$$v_w - \alpha_w \leq v \leq v_w + \alpha_w.$$

Analogously, the ground speed measurement is $v_g = v + w_g$, where $\|w_g\|_\infty \leq \alpha_g$. The LPS measurement is $p_x = p + w_x$ where p represents the actual position, and $\|w_x\|_\infty \leq \alpha_p$.

6.2.1 Sensors

In this subsection, we briefly discuss the error margin for each sensor. The error margin for those sensors plays a vital role in later sections when we address the issue of attack detection and attack correction.

Wheel speed sensor:

Wheel speed sensors are typically magnetic/optical-based sensors which provide the current wheel speed of a farming vehicle. The error margin of a wheel speed sensor varies in accordance with the sensor type, terrain conditions, tyre pressures, etc. Based on the work of [53], the error margin of the wheel speed sensors is around 5% of the current speed. Consider the worst case scenario when v takes its maximum value $v = 2.7\text{m/s}$, then $\alpha_w = 0.145\text{m/s}$.

Ground speed sensor:

Ground speed sensors are typically ultrasonic based sensors composing of two parts: a transmitter which generates an ultrasonic sound wave at 40k Hz towards the ground surface. The receiver then receives the echo from the ground. The frequency difference between the transmitted signal and the received signal is proportional to the vehicle speed [45]. The error margin of a ground speed sensor in the field test is no more than $\alpha_g = 0.027\text{m/s}$ based on the results in [35].

Local Positioning System:

A Local Positioning System (LPS) is a navigation system that provides location information in all kinds of weather, anywhere within the coverage of the network. The update

rate of a commercial LPS is around 10Hz - 20Hz, and the positioning accuracy is 'within a few millimetres' based on [66]. In this chapter, as an example, we assume the error margin for LPS is $\alpha_p = 0.005$m and the update period is $t_p = 0.1$s.

The following table summarises the error margin for the considered sensors:

Table 6.1: Sensor measurements and error margin

Sensors	Error Margin	Sensor Output Signal
Wheel speed sensor	$\alpha_w = 0.145\text{m/s}$	v_w
Ground speed sensor	$\alpha_g = 0.027\text{m/s}$	v_g
LPS	$\alpha_p = 0.005\text{m}$	p_x

6.2.2 System Model

In this section, we propose the dynamical model of the speed measurement system, first in a state-space representation, and then in an image representation. As mentioned before, for the sake of clarity, we consider a simplified system dynamics where the farming vehicle is in motion with a constant speed smaller than $v = 2.7\text{m/s}$ and at a constant direction.

We first consider the ideal speed measurement model, where we ignore any measurement noise and any attacks. In this case, since both the wheel speed sensor and the ground speed sensor provide direct speed measurements, we have

$$v = v_w = v_g,$$

where v is the actual speed signal of the farming vehicle.

As for the LPS, recall the update period is assumed to be $t_p = 0.1$; we also recall that we assume the speed is constant, and hence finite differences in position measurements gives the right speed. More specifically, we have

$$v(t) = \frac{p_x(t) - p_x(t-1)}{t_p}.$$

We can then rewrite the above equation using shift operators as follows:

$$\sigma v = \frac{(\sigma - 1)p_x}{t_p}.$$

Now we consider the system dynamics (under no measurement noise and no attacks) in a state-space representation. Since we only consider sensor attacks for this speed measurement system, for the sake of simplicity, we ignore any input signals and the state-space representation is given by:

$$\begin{cases} x(t+1) = Ax(t) \\ y(t) = Cx(t) \end{cases} \tag{6.2}$$

In this example, we choose to express the system dynamics by defining the internal latent signal $l(t) = \begin{bmatrix} v(t) \\ p(t) \end{bmatrix}$, where $v(t)$ is the actual speed of the vehicle at time t and $p(t)$ represents the actual position signal of the vehicle at time t. Vector $y(t)$ is the output signal at time t. In this example, we have $y(t) = \begin{bmatrix} v_w(t) \\ v_g(t) \\ p_x(t) \end{bmatrix}$. Since we are assuming constant speed, i.e., $\sigma v = v$, the system dynamics can be expressed as:

$$\begin{bmatrix} v(t+1) \\ p(t+1) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ t_p & 1 \end{bmatrix} \begin{bmatrix} v(t) \\ p(t) \end{bmatrix}$$
$$\begin{bmatrix} v_w(t) \\ v_g(t) \\ p_x(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} v(t) \\ p(t) \end{bmatrix}. \tag{6.3}$$

Equivalently, the ILO image representation can be expressed as

$$\begin{bmatrix} v_w \\ v_g \\ p_x \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ \sigma - 1 & 0 \\ -t_p & \sigma - 1 \end{bmatrix} \begin{bmatrix} v \\ p \end{bmatrix}. \tag{6.4}$$

Despite the fact that the input signal u is absent in this example we still call equation (6.4) an ILO image representation, as such zero input representation fits into the general concept of the ILO image representation. More details can be found in our work of [62].

6.2.3 Measurement Region

The essential difference between the situation being considered in this chapter and that of previous chapters is the influence of the noise. In previous chapters, all received signals are deterministic. Whereas in this section, the actual signal centred on the measured signal, and with a width determined by the measurement error. We now present a definition for dealing with such measurement noise, namely, the **measurement region**.

Definition 6.1. *Consider a sensor that returns a received signal x, and the error margin of this sensor is α; then, the measurement region is defined as:*

$$\mathcal{R}(\alpha, x) := \left\{ x' | x - \alpha \leq x' \leq x + \alpha \right\}. \tag{6.5}$$

Figure 6.1 illustrates the measurement region for a measurement signal x with error margin α.

Recall the assumption for bounded measurement noise, as elucidated at the beginning of this chapter. Then, the concept of measurement region can be interpreted as follows: if the sensor with error margin α produces a received signal x, then the actual signal (without measurement noise) lies within the measurement region $\mathcal{R}(\alpha, x)$.

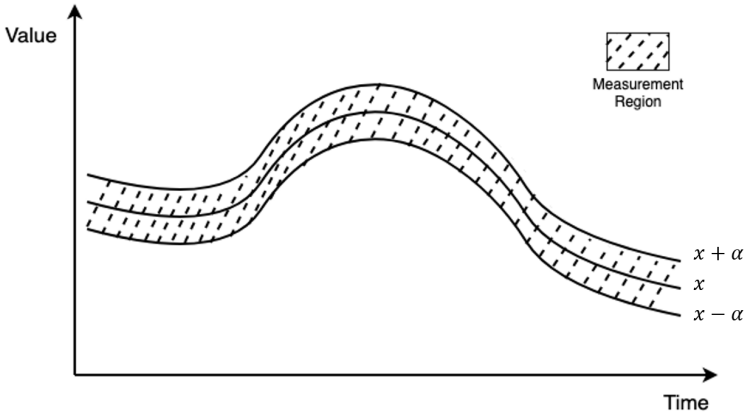

Figure 6.1: Measurement region for a signal x with accuracy of α.

6.2.4 Attack Model

We now present the system model under measurement noise and attacks. Recall the measurement noise is denoted by $w = \begin{bmatrix} w_w \\ w_g \\ w_x \end{bmatrix}$, and $\begin{bmatrix} v_w \\ v_g \\ p_x \end{bmatrix} = \begin{bmatrix} v + w_w \\ v + w_g \\ p + w_x \end{bmatrix}$. Recall that the beginning of this chapter, we are given the prior sensor knowledge that the position sensor is guaranteed to be attack-free; then, we denote the additive attack signal as $\eta = \begin{bmatrix} \eta_w \\ \eta_g \\ 0 \end{bmatrix}$. Further, we denote the attacked received signal as $r = \begin{bmatrix} r_w \\ r_g \\ r_x \end{bmatrix} = \begin{bmatrix} v_w \\ v_g \\ p_x \end{bmatrix} + \begin{bmatrix} \eta_w \\ \eta_g \\ 0 \end{bmatrix}$. Subsequently, the attacked system in ILO image representation can be expressed as:

$$\begin{bmatrix} v_w - w_w \\ v_g - w_g \\ p_x - w_x \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ \sigma - 1 & 0 \\ -t_p & \sigma - 1 \end{bmatrix} \begin{bmatrix} v \\ p \end{bmatrix} \tag{6.6}$$

$$\begin{bmatrix} r_w \\ r_g \\ r_x \end{bmatrix} = \begin{bmatrix} v_w + \eta_w \\ v_g + \eta_g \\ p_x \end{bmatrix}. \tag{6.7}$$

Equation (6.6) can be interpreted as the attack-free system dynamics under measurement noise. Such representation shows similarities as seen in the ILO image representation (equation (3.22)). The difference is that Equation (6.6) is bereft of inputs and takes into account the bounded noise.

6.3 Detection and Correction

To address the problem of attack detection and correction, we first recall the concept of the system behaviour $\mathcal{B}$. In this case, we define the behaviour of the system via an ideal image representation where we ignore any measurement noise and attack signals, after which the behaviour $\mathcal{B}$ of the system is the set given by

$$\mathcal{B} =: \left\{ y = \begin{bmatrix} v_w \\ v_g \\ p_x \end{bmatrix} : \mathbb{Z}_+ \to \mathbb{R}^3 \middle| \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ \sigma - 1 & 0 \\ -t_p & \sigma - 1 \end{bmatrix} \begin{bmatrix} v \\ p \end{bmatrix}, \text{ for some } v, p \right\}.$$

Since we are considering prior sensor knowledge on p_x, we then recall Equation (3.36), the behaviour of interest in this example is as follows:

$$\mathcal{B}_{hom-ps} : \left\{ \begin{bmatrix} v_w \\ v_g \\ 0 \end{bmatrix} \middle| \begin{bmatrix} v_w \\ v_g \\ 0 \end{bmatrix} \in \mathcal{B} \right\}. \tag{6.8}$$

Based on Theorem 3.14, it is evident that given prior sensor knowledge where the positioning sensor is attack-free, the polynomial matrix $\begin{bmatrix} 0 & 1 \\ \xi - 1 & 0 \\ -t_p & \xi - 1 \end{bmatrix}$ is left unimodular. This, in turn, implies that the system is trivially secure, i.e., $\delta_{s-ps}(\Sigma) = 3$. The implication is that the system can detect/correct any attacks.

Now let us consider what will happen if we do not have the prior knowledge on

the positioning sensor. More specifically, let us consider a case when the attacker can potentially attack any sensor measurements. In this case, the behaviour of interest is the entire $\mathcal{B}$. Recall the definition of sensor security index $\delta_s(\Sigma) := \min_{0 \neq y \in \mathcal{B}} \|y\|$ as in Definition 3.8. Using Theorem 3.8, the sensor security index $\delta_s(\Sigma) = 1$. From another perspective, the sparsest non-zero signals in $\mathcal{B}$ has weight 1, for example, a constant signal $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \in \mathcal{B}$ has weight 1. In this case, detection and correction cannot be guaranteed even if only one sensor is attacked.

Such observation confirms our previous statement in Chapter 3. The vulnerability of the system with prior sensor knowledge is no worse than the same system without such prior sensor knowledge. In this case, the original system is not trivially secure and has a small security index. When we provide certain prior sensor knowledge on one of the sensors, the system is then trivially secure and as a result, can detect/correct any attacks.

In this particulate example, we can make sense of such difference between $\delta_{s-ps}(\Sigma)$ and $\delta_s(\Sigma)$. Without the prior sensor knowledge, the attacker can simply attack the position sensor by altering the value of the signal whilst maintaining the 'slope' (or gradient) of the attack-free signal as before. Then, the user cannot detect/correct this particular type of attack. However, if we know that the position sensor is guaranteed to be attack-free, it then becomes possible to reconstruct the true speed signal in the noise-case and provide a bounded region that contains the true speed signal in the noisy case. This is common sense but still illustrates the value of the proposed theory in the thesis.

We now seek methods to apply our previously proposed attack detection and attack correction methods to this particular speed measurement system where the measurement noise is being considered.

Recall that the system is given by equation (6.6), which denotes the attack-free system under measurement noise. The wheel speed sensor measurement v_w and the ground speed sensor measurement v_g provide 'direct' measurements of the speed signal v. More specifically, we take the wheel speed sensor as an example: given the error margin α_w

and the measurement signal v_w; the following equation holds:

$$v \in \mathcal{R}(\alpha_w, v_w). \tag{6.9}$$

A similar analysis can be applied to the ground speed sensor, and we have

$$v \in \mathcal{R}(\alpha_g, v_g). \tag{6.10}$$

The situation for the Local Positioning System is slightly different. Instead of directly measuring the latent signal, we have $\frac{\sigma-1}{t_p}(p_x - w_x) = v$. Based on this we can see that for the position measurement, the noise signal is $\frac{w_x - \sigma w_x}{t_p}$. As a result, the error margin for such a noise signal will change. The shift operator σ will double the error margin. The sampling period t_p will also increases (or decreases if $t_p > 1$) the error margin by factor of $\frac{1}{t_p}$. To summarise, the error margin of the resulting noise for the LPS is $\alpha'_p = \frac{2}{t_p}\alpha_p$.

To address the problem of attack detection and correction, the following three measurement regions are of particular interest:

$$\begin{aligned} \mathcal{R}_w &= \mathcal{R}(\alpha_w, r_w) \\ \mathcal{R}_g &= \mathcal{R}(\alpha_g, r_g) \\ \mathcal{R}_p &= \mathcal{R}(\frac{2}{t_p}\alpha_p, \frac{\sigma-1}{t_p}r_x). \end{aligned} \tag{6.11}$$

6.3.1 Attack Detection

We now propose attack detection Algorithm 12 for this specific system given prior knowledge that the positioning sensor is guaranteed to be attack-free:

Algorithm 12 Sensor attack detection for farming vehicle speed measurement system

Construct the measurement region $\mathcal{R}_w, \mathcal{R}_g, \mathcal{R}_p$ as in equation (6.11)
Calculate $S = \mathcal{R}_w \cap \mathcal{R}_g \cap \mathcal{R}_p$
if $S \neq \emptyset$ **then** decide no attack.
else decide attack occurred.
end if

When comparing with our previous sensor attack detection Algorithm 3, it is evident that both the algorithms determine whether there exists an attack signal based on a specific indicator. In Algorithm 3, a determined residual signal s is used, while in Algorithm 12, a signal set S is used due to the existence of measurement noise.

Before presenting the next proposition regarding the performance of Algorithm 12, we define $\alpha_{max} = \max(\alpha_w, \alpha_g, \frac{2}{t_p}\alpha_p)$, where $\max(\bullet)$ returns the largest signal in $\bullet$.

Proposition 6.1. *Algorithm 12 returns the correct detection result if the attack signal* $\eta = \begin{bmatrix} \eta_w \\ \eta_g \\ 0 \end{bmatrix}$ *satisfies the following assumption:*

The l_∞ *norm for each of the sensor attack signals satisfies:*

$$\|\eta_w\|_\infty > 2\alpha_{max}\|v_w\|_\infty$$
$$\|\eta_g\|_\infty > 2\alpha_{max}\|v_g\|_\infty.$$

Proof. First note that when no attack occurs, i.e., $\eta = 0$; then, we have $v \in \mathcal{R}_w, v \in \mathcal{R}_g$ and $v \in \mathcal{R}_p$ and thus, $S \neq \emptyset$ must hold.

When an attack presents and satisfies the assumption, then at least one sensor is attacked and at least one sensor is attack-free. If we choose one attacked sensor and one attack-free sensor, then their corresponding measurement regions have no intersection, and thus, $S = \emptyset$. □

The assumption in proposition 6.1 means that if the attack signal is significant enough, then the detection algorithm can detect such an attack. On the other hand, if the maximum value of the attack signal (l_∞ norm of the attack signal) is smaller than the 'width' of the largest measurement region, then it is possible to be undetectable for such an attack signal. However, the consequence of such undetectable attacks is no worse than the influence of the bounded noise. The risk of such attacks should already be managed based on the noise model. In other words, such undetectable attacks may not be significant enough to cause severe damage to the system within a short period of time.

6.3.2 Attack Correction

Our objective is to reconstruct the speed signal of the farming vehicle based on the prior knowledge (as mentioned in (6.1.4)) that the positioning sensor is guaranteed to be attack-free. Recall previous chapters, where the essential step towards attack correction is to find the observers for reconstructing the latent signal. Recall the ideal system model (without noise and attack) which satisfies the following equation:

$$\begin{bmatrix} y \\ 0 \end{bmatrix} = \begin{bmatrix} M(\sigma) \\ D(\sigma) \end{bmatrix} \begin{bmatrix} v \\ p \end{bmatrix} = \left[\begin{array}{cc} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ \hline \sigma - 1 & 0 \\ -t_p & \sigma - 1 \end{array}\right] \begin{bmatrix} v \\ p \end{bmatrix}. \tag{6.12}$$

Since the system is trivially secure when providing the prior sensor knowledge, the submatrix $\begin{bmatrix} 0 & 1 \\ \sigma - 1 & 0 \\ -t_p & \sigma - 1 \end{bmatrix}$ is left unimodular. This means that in order to reconstruct the speed signal or achieve latent signal estimation, we only need one observer. In other words, when the system is under sensor attack and the noise is ignored, we can achieve attack correction using only the attack-free position measurement signal p_x (=r_x).

The observer for attack correction is formed by choosing among all possible polynomial left inverse of $\begin{bmatrix} 0 & 1 \\ \sigma - 1 & 0 \\ -t_p & \sigma - 1 \end{bmatrix}$. An observer for this simple example is easily found to be equation (6.13)

$$\begin{bmatrix} M^{(3,\bullet)}(\xi) \\ D(\xi) \end{bmatrix}^{-1} = \begin{bmatrix} \frac{\xi - 1}{t_p} & 0 & -\frac{1}{t_p} \\ 1 & 0 & 0 \end{bmatrix} \tag{6.13}$$

In the presence of noise and possible attacks, the received signal $r = \begin{bmatrix} r_w \\ r_g \\ r_x \end{bmatrix}$. We now propose the following equation to reconstruct the latent signal:

$$\begin{bmatrix} \hat{v} \\ \hat{p} \end{bmatrix} = \begin{bmatrix} \frac{\sigma-1}{t_p} & 0 & -\frac{1}{t_p} \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} r_x \\ 0 \\ 0 \end{bmatrix}. \tag{6.14}$$

After the latent signal is being reconstructed, we can then read-off the estimated speed signal $\hat{v}$ from the resulting latent signal. It is evident that $v \in \mathcal{R}(\frac{2}{t_p}\alpha_p, \hat{v})$ must hold, thus meaning that the estimated speed signal is close to the actual speed of the farming vehicle.

6.4 Further Discussion

In the previous discussions in this chapter, we assumed a certain prior sensor knowledge, that is, we assumed that the positioning sensor is guaranteed to be attack-free. We also explained that if we do not consider such prior sensor knowledge, the system is more vulnerable to attacks, in the sense that the sensor security index is smaller without such prior sensor knowledge. In this case, even when the attacker can attack only one sensor, there is no guarantee to achieve attack correction.

However, in many engineering applications, we are not interested in reconstructing all the sensor measurements. Put differently, a subset of measurements may be of greater interest compared with the rest of the measurements. For example, in this chapter, we state that our top priority is to reconstruct the speed signal while reconstructing the position signal is not of primary interest. As mentioned in the beginning in this chapter, the speed of the vehicle is an important safety parameter. Taking this into account, we can achieve speed signal attack correction when a single sensor is attacked.

Consider the case when the system is under single sensor attack but we do not know which sensor is attacked. Under the assumption that the attack signal is significant enough as seen in Proposition 6.1, we first calculate $S_1 = \mathcal{R}_w \cap \mathcal{R}_g$. If $S_1 \neq \emptyset$, it then implies that the two speed sensors are both attack-free and that the actual speed signal v satisfies $v \in S_1$. If $S_1 = \emptyset$, then we immediately know that one of the speed sensors is attacked and that the position sensor must be attack-free. In this case, we have $v \in \mathcal{R}_p$.

In both cases ($S_1 = \emptyset$ or $S_1 \neq \emptyset$), reconstructing the speed signal is possible.

To summarise our observations above, if we are only interested in reconstructing a subset of sensors instead of all sensors, then it is possible to relax the detectability and correctability constraints regarding the number of sensors being attacked. This is an interesting topic because there is a hierarchy of the sensor measurement signals. Speed can be obtained from position but not vice-versa. This proposition in this chapter exhibits similarities with unequal error control code and selectable detection and correction levels [49]. Developing novel results towards a new topic, namely, unequal attack correction, which is based on different correction levels or correction priorities regarding CPS security, these are possible future research directions followed by this thesis.

6.5 Comparison and Conclusion

The problem of attack detection and attack correction with noisy measurement has been discussed before in the literature. For example, in [41], state estimation methods against sensor attack have been discussed in the presence of Gaussian measurement noise. The authors in [41] also assume an upper bound on the number of sensor attacks. Attack detection is discussed using a Kalman filter-based approach in [41]. Since a Gaussian measurement noise is considered in [41], an 'optimum guarantee' on the achievable state estimation error is discussed. As mentioned in Chapter 2.1.5, the notion of 'optimum guarantee' is close to but never equivalent to 'guarantee'. In this chapter, we consider bounded noise instead of Gaussian noise. We do not choose to model the noise as Gaussian noise, because our aim is to produce a guaranteed detection/correction algorithm. Nevertheless, solving the CPS security issue for systems under Gaussian noise using the notion of behavioural approach observer-based correction algorithm are of interest for us in the context of future research.

This chapter is also related to the work of [27], in which the estimation problem of a sparse signal under bounded noise is considered. There is alignment of our result in this chapter with [27]. In both cases, detection is possible if the noise is not significant compared with the measurements. It is noteworthy that [27] uses linear equations with

no system dynamics, while in our case, a specific system setup is considered in order to illustrate the potential research directions for this thesis.

The work [56] focuses on sensor attack observer design under bounded noise using a Luenberger-liked state observer in a more general setting. In [56], the estimation error is guaranteed to be bounded (with a known upper bound). The upper bound is stated in terms of the eigenvalues of the system matrix A. Meanwhile in our case, the error bound is also related to the eigenvalues of A (that is why we have $1/t_p$ in the error margin). Moreover, all the problem settings and results in this chapter can be potentially applied to other types of attacks (actuator attack, for example), using the previously proposed techniques in this thesis. Whereas in [56], problems such as actuator attack correction are not being mentioned.

From the simple engineering example mentioned in this chapter, we realise that it is possible to address noisy attack detection and correction using the notion of set-membership approach. The set-membership approach is appropriate if there is limited stochastic information about the measurement uncertainty or the measurement uncertainty cannot be modelled in a probabilistic way. Such property coincides with our behavioural approach and observer-based correction algorithm. The essence for the set-membership approach, as discussed in, for example, [31, 32, 54, 64] is to update the state estimation for a dynamical system recursively. Even though the example proposed in this chapter is simple, it represents a first step to extend our approach to the noisy case.

To conclude this chapter, we can see that the notion of security index remained powerful when we discuss the security problems under the noisy case. Moreover, it is only when the attacks are detectable/correctable in the noise-free case that we can expect to achieve detection/correction in the bounded noise case. Nevertheless, we need to reiterate that the contents in this chapter are still in the nascent stages of research. This is our first step to apply the concept of the system's security index, and corresponding detection/correction algorithms to a noisy system. Future works include different movement patterns (non-constant speed, accelerating/decelerating, changing directions), more complex system models (more sensors and actuators) and different attack models (actuator attacks in the noisy case).

Chapter 7
Conclusion and Future Research Directions

7.1 Conclusion

In this thesis, we addressed the problem of attack detection and correction for MIMO LTI CPS under various types of attacks.

In Chapter 1, we briefly explained the background of CPS security. We also presented our motivation of this thesis. In Chapter 2, we reviewed the relevant literature in the area of CPS security, along with some classic fault detection and identification literature.

In Chapter 3, we presented the vulnerability analysis for a system under sensor only attacks. Sufficient conditions for successful sensor attack detection and correction were stated using the notion of the sensor security index. Two different attack detection algorithms were proposed in this chapter, one starting from a kernel representation, whereas the other started from an ILO image representation. Both detection algorithms can be used as universal algorithms in order to detect different types of attack signals being considered in subsequent chapters. Attack correction methods for sensor only attack was also presented in this chapter.

In Chapter 4, we extended the concept of security index to apply to systems under actuator only attacks, namely, the actuator security index. Attack correction algorithm was proposed for dealing with such actuator only attacks.

In Chapter 5, under the notion of strong observability assumption, we proposed the concept of a correction index to address the problem of attack correction for system under both actuator and sensor attacks.

In Chapter 6, we illustrated the working of the proposed detection and correction algorithm for a specific speed measurement system where the measurements are subject to bounded noise.

7.2 Summary of Contributions

• Security index for a variety of attack models

In this thesis, we analysed the vulnerability of CPSs against adversarial attacks using the notion of security index. We extended the concept of the security index to make this notion applicable for various attack scenarios. For each attack scenarios, sufficient conditions for attack detectability and correctability are stated using the corresponding security index. For each attack scenarios, the corresponding security index is stated using a representation-free manner which indicating that the notion of security index is a true system property, unlike the notion of say controllability, which depends on the representation. Followed by such representation-free system property, we then characterise the system vulnerability for different representations (e.g., kernel or ILO image representation).

• CPS attack detection and correction method

For systems under adversarial attacks, two questions are particularly noteworthy: attack detection and attack correction. In this thesis, we addressed the problem of guaranteed attack detection and correction for various types of attack.

For attack detection, we proposed two universal algorithms that can be applied to any attack scenario. The proposed detection algorithms were guaranteed to produce the correct outcome under the assumption that the number of sensors/actuators being attacked is upper bounded by a specific value.

In terms of attack correction, we proposed corresponding correction algorithms for different attack scenarios. The correction algorithms were guaranteed to produce the correct attack signal under the assumption that an upper bound on the number of attacked sensor and/or actuators is given.

Moreover, the correction algorithm uses less observers as compared to e.g., [13, 56].

Both the detection and correction algorithms are intuitive and easier to understand when compared to the optimisation-based estimation method as mentioned in [22, 57]. The overall computational complexity is significantly smaller than the optimisation-based approach since all the observers can be pre-computed.

• **Prior attack knowledge**

In addition to different attack scenarios, we further considered a certain level of prior attack knowledge. In the thesis such an assumption concerns prior knowledge about a subset of sensors and/or actuators that are attack-free. Based on such prior knowledge, we performed a vulnerability analysis and proposed attack detection as well as correction algorithms. It is expected that the security index changes when we are provide with such prior knowledge. More specifically, the security index subject to a specific prior knowledge is no worth than that without the prior knowledge. In particular, if there is sufficient information to achieve attack detection/correction with only the attack-free sensors/actuators, then such systems are trivially secure and can detect/correct any attacks. From another point of view, redundancy in measurement/actuation is important for CPS security.

7.3 Suggested Future research

Future research can be divided into the following categories.

Generalise beyond deterministic linear time invariant systems

In this thesis, we have illustrated the working of the proposed attack detection and correction method using numerical examples and a specific speed measurement system as an example. However, developing theories to address the problem of attack detection and attack correction for other system models are still under development and are the top priority for our future research. The models of interest include systems with unbounded disturbance and/or noise, time-varying systems etc. Behaviour theorem for non-linear system security is in its infancy so any generalisation may be difficult, instead, considering the security issue for finite word-length systems using behavioural approach is realistic and will be the focus of our future research.

Minimal delay detection/correction

In this thesis, certain extent of delay exists for both detection and correction algorithms. Generally speaking, such delays are inevitable. However, it is possible to develop theorems and concepts around minimising the amount of delay. More specifically, how to choose the optimum observers or how to choose a parity relation so that detection/correction can be achieved with minimum delay in discrete time remains an open research problem. The concept of minimal lag systems in the behaviour theory might be a good starting point regarding this problem. Moreover, the work of [26] discussed the minimum basis in the transfer function approach, which is also likely to serve as a starting point when we want to solve the delay issue using polynomial matrices and a behavioural approach.

Algorithm implementation

In this thesis, we proposed novel approaches to address the problem of attack detection and attack correction for linear CPS. However, designing an engineering system in a suitable engineering application in order to verify those novel ideas can be an onerous task. For example, how likely a real-world engineering system has a high security index? How to compute the error margin when the system model become more complex (compare with our example in Chapter 6). We are curious to observe the performance of our proposed concepts and algorithms in real-world engineering applications.

Bibliography

[1] "Agri view: Driverless tractors," http://www.merlofarminggroup.com/agri-view-driverless-tractors, accessed: 2019-09-21.

[2] A. Adewoyin, "Fuel consumption evaluation of some commonly used farm tractors for ploughing operations on the sandy-loam soil of oyo state, nigeria," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 6, pp. 2865–2871,, 06 2013.

[3] N. Ahmed, D. De, and I. Hussain, "Internet of things (iot) for smart precision agriculture and farming in rural areas," *IEEE Internet of Things Journal*, vol. 5, no. 6, pp. 4890–4899, Dec 2018.

[4] J. H. Ahrens and H. K. Khalil, "High-gain observers in the presence of measurement noise: A switched-gain approach," *Automatica*, vol. 45, no. 4, pp. 936 – 943, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0005109808005591

[5] T. Alamo, J. Bravo, and E. Camacho, "Guaranteed state estimation by zonotopes," *Automatica*, vol. 41, no. 6, pp. 1035 – 1043, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0005109805000294

[6] R. Baheti and H. Gill, "Cyber-physcal systems," *The Impact of Control Technology*, 2011.

[7] S. BEALE and B. SHAFAI, "Robust control system design with a proportional integral observer," *International Journal of Control*, vol. 50, no. 1, pp. 97–111, 1989. [Online]. Available: https://doi.org/10.1080/00207178908953350

[8] M. Blanke, M. Kinnaert, J. Lunze, and M. Staroswiecki, *Diagnosis and Fault-Tolerant Control*, 3rd ed. Springer Publishing Company, Incorporated, 2015.

[9] E. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, pp. 4203–4215, 2005.

[10] J. Chen and R. J. Patton, *Robust Model-Based Fault Diagnosis for Dynamic Systems*. Kluwer, 1999.

[11] J. Chen, R. Patton, and H.Zhang, "Design of unknown input observers and robust fault detection filters," *International Journal of Control*, vol. 63, no. 1, pp. 85–105, 1996.

[12] Y. Chen, S. Kar, and J. Moura, "Cyber-physical systems: dynamic sensor attacks and strong observability," in *Proc. 40th International Conference on Acoustics, Speech and Signal Processing*, Brisbane, Australia, 2015, pp. 1752–1756.

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

[14] M. Chong and M. Kuijper, "Characterising the vulnerability of linear control systems under sensor attacks using a system' s security index," in *Proc. IEEE 55th Conference on Decision and Control*, Las Vegas, USA, December, 2016, pp. 5906–5911.

[15] ——, "Vulnerability of linear systems against sensor attacks–a system's security index," in *Proc. 22nd International Symposium on Mathematical Theory of Networks and Systems*, Minneapolis, USA, July, 2016, pp. 373–376.

[16] E. Chow and A. Willsky, "Analytical redundancy and the design of robust failure-detection systems," *IEEE Transactions on Automatic Control*, vol. 29, pp. 603–614, 1984.

[17] E. Y. Chow, "A failure detection system design methodology," Ph.D. dissertation, Massachusetts Institute of Technology, 1980.

[18] F. Colonius, U. Helmke, D. Prätzel-Wolters, and F. Wirth, *System and Control: Foundations and Applications*. Springer, 2001.

[19] CRS, "Congressional research service report," 2011. [Online]. Available: https://http://www.loc.gov/crsinfo/about/crs11_annrpt.pdf

[20] S. X. Ding, T. Jeinsch, P. M. Frank, and E. L. Ding, "A unfined approach to the optimization of fault detection systems," *International Journal of Adaptive Control and Signal Processing*, vol. 14, pp. 725–745, 2000.

[21] X. Ding, L. Guo, and T. Jeinsch, "A characterization of parity space and its application to robust fault detection," *IEEE Transactions on Automatic Control*, vol. 44, no. 2, pp. 337–343, 1999.

[22] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Transactions of Automatic Control*, vol. 59, no. 6, pp. 1454–1467, 2014.

[23] E. Fogel and Y. Huang, "On the value of information in system identification—bounded noise case," *Automatica*, vol. 18, no. 2, pp. 229 – 238, 1982. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0005109882901108

[24] E. Frisk, "Residual generation for fault diagnosis: Nominal and robust design," Ph.D. dissertation, Link"oping University, 1998.

[25] E. Frisk and M. Nyberg, "Using minimal polynomial bases for fault diagnosis," in *Proc. European Control Conference*, Karlsruhe, Germany, 1999, pp. 4161 – 4166.

[26] ——, "A minimal polynomial basis solution to residual generation for fault diagnosis in linear systems," *Automatica*, vol. 37, no. 9, pp. 1417–1424, 2001.

[27] J. J. Fuchs, "Recovery of exact sparse representations in the presence of bounded noise," *IEEE Transactions on Information Theory*, vol. 51, no. 10, pp. 3601–3608, Oct 2005.

[28] P. A. Fuhrmann and U. Helmke, *The mathematics of networks of linear systems*. Springer, 2015.

[29] J. Gertler, "Fault detection and isolation using parity relations," *Control Engineering Practice*, vol. 5, no. 5, pp. 653 – 661, 1997. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0967066197000476

[30] J. M. Hendrickx, K. H. Johansson, R. M. Jungers, H. Sandberg, and K. C. Sou, "Efficient computations of a security index for false data attacks in power networks," *IEEE Transactions on Automatic Control*, vol. 59, no. 12, pp. 3194–3208, 2014.

[31] R. Hill and Y. Luo, "Exact recursive updating of uncertainty sets for discrete-time plants with a lag," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, Dec 2017, pp. 971–976.

[32] R. Hill, Y. Luo, and U. Schwerdtfeger, "Exact recursive updating of state uncertainty sets for linear siso systems," *Automatica*, vol. 95, pp. 33 – 43, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0005109818302528

[33] D. Hinrichsen and D. Prätzel-Wolters, "Generalized Hermite matrices and complete invariants of stict system equivalence," *SlAM J. Control and Optimization*, vol. 21, pp. 289–306, 1983.

[34] I. Hwang, S. Kim, Y. Kim, and C. E. Seah, "A survey of fault detection, isolation, and reconfiguration methods," *IEEE Transactions on Control System Technology*, vol. 18, no. 3, pp. 636–652, 2010.

[35] K. Imou, M. Ishida, T. Okamoto, Y. Kaizu, A. Sawamura, and N. Sumida, "Ultrasonic doppler sensor for measuring vehicle speed in forward and reverse motions including low speed motions," *Agricultural Engineering International: the CIGR Journal of Scientific Research and Development*, vol. 3, 2001.

[36] X. Jin, W. M. Haddad, and T. Yucelen, "An adaptive control architecture for mitigating sensor and actuator attacks in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 62, no. 11, pp. 6058–6064, Nov 2017.

[37] R. E. Kalman, "On the general theory of control systems," Moscow, 1960, pp. 481–492.

[38] M. Kuijper, *First-Order Representations of Linear Systems*. Boston, USA: Birkhäuser, 1994.

[39] M. Li, K. Imou, K. Wakabayashi, and S. Yokoyama, "Review of research on agricultural vehicle autonomous guidance," *International Journal of Agricultural and Biological Engineering*, vol. 2, 09 2009.

[40] R. K. Mehra and J. Peschon, "An innovations approach to fault deection and diagnosis in dynamic systems," *Automatica*, vol. 7, pp. 637–640, 1971.

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

[42] S. Mishra, Y. Shoukry, N. Karamchandani, S. N. Diggavi, and P. Tabuada, "Secure state estimation against sensor attacks in the presence of noise," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 1, pp. 49–59, March 2017.

[43] S. Mitra, N. Seifert, M. Zhang, Q. Shi, and K. S. Kim, "Robust system design with built-in soft-error resilience," *Computer*, vol. 38, no. 2, pp. 43–52, Feb 2005.

[44] M. Pajic, P. Tabuada, I. Lee, and G. J. Pappas, "Attack-resilient state estimation in the presence of noise," in *2015 54th IEEE Conference on Decision and Control (CDC)*, Dec 2015, pp. 5827–5832.

[45] *True Ground Speed Sensor (TGSS)*, Parker Hannifin Corporation, 2012.

[46] J. Polderman and J. Willems, *Introduction to mathematical systems theory: a behavioral approach*. Springer, 1997, vol. 26.

[47] J. Proakis and M. Salehi, *Digital Communications*. McGraw Hill, 2008.

[48] S. Roy and R. A. Iltis, "Decentralized linear estimation in correlated measurement noise," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 27, no. 6, pp. 939–941, Nov 1991.

[49] L.-J. Saiz-Adalid, P.-J. Gil-Vicente, J.-C. Ruiz-García, D. Gil-Tomás, J. C. Baraza, and J. Gracia-Morán, "Flexible unequal error control codes with selectable error detection and correction levels," in *Computer Safety, Reliability, and Security*, F. Bitsch, J. Guiochet, and M. Kaâniche, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 178–189.

[50] H. Sandberg, A. Teixeira, and K. H. Johansson, "On security indices for state estimators in power networks," in *First workshop on secure control system*, 2010, pp. 1 – 6.

[51] H. Sandberg and A. M. H. Teixeira, "From control system security indices to attack identifiability," in *Proc. 2016 Science of Security for Cyber-Physical Systems Workshop (SOSCYPS)*, 2016, pp. 1–6.

[52] D. Sauter and F. Hamelin, "Frequency-domain optimization for robust fault detection and isolation in dynamic systems," *IEEE Transactions on Automatic Control*, vol. 44, no. 4, pp. 878–882, 1999.

[53] R. Schwarz, O. Nelles, P. Scheerer, and R. Isermann, "Increasing signal accuracy of automotive wheel-speed sensors by online learning," in *Proceedings of the 1997 American Control Conference (Cat. No.97CH36041)*, vol. 2, June 1997, pp. 1131–1135 vol.2.

[54] F. Schweppe, "Recursive state estimation: Unknown but bounded errors and system inputs," *IEEE Transactions on Automatic Control*, vol. 13, no. 1, pp. 22–28, February 1968.

[55] S. Seshu and R. Waxman, "Fault isolation in conventional linear system – a feasibility study," *IEEE Transaction on Reliability*, vol. 15, pp. 11–16, 1966.

[56] Y. Shoukry, M. Chong, M. Wakaiki, P. Nuzzo, A. L. Sangiovanni-Vincentelli, S. A. Seshia, J. P. Hespanha, and P. Tabuada, "Smt-based observer design for cyber-physical systems under sensor attacks," in *2016 ACM/IEEE 7th International Conference on Cyber-Physical Systems (ICCPS)*, April 2016, pp. 1–10.

[57] M. Showkatbakhsh, Y. Shoukry, R. H. Chen, S. Diggavi, and P. Tabuada, "An smt-based approach to secure state estimation under sensor and actuator attacks," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, 2017, pp. 157–162.

[58] D. Silvestre, P. Rosa, J. P. Hespanha, and C. Silvestre, "Fault detection for lpv systems using set-valued observers: A coprime factorization approach," *Systems and Control Letters*, vol. 106, pp. 32 – 39, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167691117301019

[59] J. Slay and M. Miller, *Critical infrastructure protection*. Springer, 2007.

[60] T. Song and E. G. C. Jr, "Robust H_2 estimation with application to robust fault detection," *Journal of Guidance*, vol. 23, no. 6, pp. 1067–1071, 2000.

[61] Z. Tang, M. Kuijper, M. Chong, I. Mareels, and C. Leckie, "Sensor attack correction for linear systems with known inputs," *IFAC-PapersOnLine*, vol. 51, no. 23, pp. 206 – 211, 2018, 7th IFAC Workshop on Distributed Estimation and Control in Networked Systems NECSYS 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2405896318335687

[62] Z. Tang, M. Kuijper, M. S. Chong, I. Mareels, and C. Leckie, "Linear system security - detection and correction of adversarial sensor attacks in the noise-free case," *Automatica*, vol. 101, pp. 53–59, 2019. [Online]. Available: https://doi.org/10.1016/j.automatica.2018.11.048

[63] ——, "Attack correction for noise-free linear systems subject to sensor attacks," *23rd International Symposium on Mathematical Theory of Networks and Systems (MTNS)*, pp. 18–21, July, 2017.

[64] R. Tempo, "Robust estimation and filtering in the presence of bounded noise," *IEEE Transactions on Automatic Control*, vol. 33, no. 9, pp. 864–867, Sep. 1988.

[65] A. M. Tillmann and M. E. Pfetsch, "Eyclidean sectionss of l_1^n with sublinear randomness and error-correction over the reals," in *Proc. RANDOM 08*, 2008, pp. 444–454.

[66] TOPCON, "Local positioning system," 2019, original document from TopconCare. [Online]. Available: https://topconcare.com/en/hardware/mc-sensors/lps-local-positioning-system/specifications

[67] J. S. Warner and R. G. Johnston, "A simple demonstration that the global positioning system is vulnerable to spoofing," *Journal of Security Administration*, pp. 19–27, 2002.

[68] K. Watanabe and D. M. Himmelblau, "Instrument fault detection in systems with uncertainties," *International Journal of Systems Science*, vol. 13, no. 2, pp. 137–158, 1982.

[69] Weitian Chen and M. Saif, "Fault detection and isolation based on novel unknown input observer design," in *2006 American Control Conference*, June 2006, pp. 6 pp.–.

[70] A. W. Werth, "Towards distinguishing between cyber-attacks and faults in cyber-physical systems," Master's thesis, Vanderbilt University, 2014.

[71] J. C. Willems, Ed., *Models for dynamics*. Dynamics Reported, 1989.

[72] J. C. Willems, "Paradigms and puzzles in the theory of dynamical systems," *IEEE Transactions on Automatic Control*, vol. 36, no. 3, pp. 259–294, March 1991.

[73] H. Ye, G. Wang, and D. S, "A new parity space approach for fault detection based on stationary wavelet transform," *IEEE Transactions on Automatic Control*, vol. 49, no. 2, pp. 281–287, 2004.