SECURITY OF LINEAR CONTROL SYSTEMS

A Dissertation

by

BHARADWAJ SATCHIDANANDAN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	P. R. Kumar
Committee Members,	Riccardo Bettati
	I-Hong Hou
	Dileep Kalathil
	Srinivas Shakkottai
	Le Xie
Head of Department,	Miroslav M. Begovic

August 2019

Major Subject: Electrical Engineering

ABSTRACT

The coming decades may see the large scale deployment of networked cyber-physical systems to address global needs in areas such as energy, water, healthcare, and transportation. However, as recent events have shown, such systems are vulnerable to cyber attacks. They are not only econoically important, but being safety critical, their disruption or misbehavior can also cause injuries and loss of life. It is therefore important to secure such networked cyber-physical systems against attacks. In the absence of credible security guarantees, there will be resistance to the proliferation of cyber-physical systems, which are much needed to meet global needs in critical infrastructures and services.

This study addresses the problem of secure control of networked cyber-physical systems. This problem is different from the problem of securing the communication network, since cyber-physical systems at their very essence need sensors and actuators that interface with the physical plant, and malicious agents may tamper with sensors or actuators, as recent attacks have shown.

We consider physical plants that are being controlled by multiple actuators and sensors communicating over a network, where some sensors and actuators could be “malicious." A malicious sensor may not report the measurement that it observes truthfully, while a malicious actuator may not apply actuation signals in accordance with the designed control policy.

In the first part of this work, we introduce, against this backdrop, the notions of securable and unsecurable subspaces of a linear dynamical system, and show that they have important operational meanings for both deterministic and stochastic linear dynamical systems in the context of secure control. These subspaces may be regarded as analogs of the controllable and unobservable subspaces reexamined in an era where there is intense interest in cybersecurity of control systems.

In the second part of the work, we propose a general technique, termed “Dynamic Watermarking,” by which honest nodes in the system can detect the actions of malicious nodes, and disable closed-loop control based on their information. Dynamic Watermarking employs the technique of honest actuators injecting a “small" random noise, known as private excitation, into the system

which will reveal tampering of measurements by malicious sensors. We lay the foundations for the theory for how such an active defense can be used to secure networked systems of sensors and actuators.

To my advisor, my teachers, my parents, and my brother.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

The contents of Chapter 9 are based on joint work with Dr. Woo-Hyun Ko of Texas A&M University.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

This dissertation is based on work supported by NSF Science & Technology Center Grant CCF-0939370, NSF Contract Nos. CNS-1302182, CNS-1646449, CCF-1619085, ECCS-1547075, the US Army Research Office under Contract Nos. W911NF-15-1-0279, W911NF-18-10331, the US Army Research Laboratory under Contract No. W911NF-19-2-0033, the AFOSR under Contract No. FA-9550-13-1-0008, Office of Naval Research under Contract N00014-18-1-2048, the Power Systems Engineering Research Center (PSERC), and NPRP grant NPRP 8-1531-2-651 from the Qatar National Research Fund, a member of Qatar Foundation.

NOMENCLATURE

ARX	Autoregressive Input
ARMAX	Autoregressive Moving Average Input
CPS	Cyber-Physical System
DoS	Denial of Service
IID	Independent and Identically Distributed
LQG	Linear Quadratic Gaussian
LQR	Linear Quadratic Regulator
MIMO	Multiple-Input-Multiple-Output
MMSE	Minimum Mean-Square Error
MST	Martingale Stability Theorem
SISO	Single-Input-Single-Output

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

1. INTRODUCTION*

The 21^{st} century could well be the era of large-scale system building. Several large-scale cyber-systems are envisioned to be formed by the interconnection of many embedded devices communicating with each other, and interacting with the physical world. They include smart energy grid, intelligent transportation systems, internet of things, telesurgical systems, distributed chemical plants, and robotics. Their operation requires tight integration of communication, control, and computation, and they have been broadly termed as Cyber-Physical Systems (CPS).

While the importance and benefits of cyber-physical systems require no emphasis, their sustained proliferation is contingent on some key challenges being addressed, security being a primary one. Since CPSs have many applications in safety-critical scenarios, security breaches of these systems can have adverse consequences including economic loss, injury and death.

There have been many instances of demonstrated attacks on cyber-physical systems in the recent past [3, 4]. In Maroochy-Shire, Australia, in the year 2003, a disgruntled ex-employee of a sewage treatment corporation hacked into the computers controlling the sewage system and issued commands which led to a series of faults in the system [5, 3]. This is an insider attack, where the adversary has the necessary credentials to access and issue control commands to the system. We will return to this point later. Another example is the attack on computers controlling the Davis-Besse nuclear power plant in Ohio. The Slammer worm, which infected about $75,000$ hosts in the internet in under ten minutes, also infected the computers controlling the nuclear power plant, disabling the safety monitoring systems [3]. While the Slammer worm was not designed to target the nuclear power plant, the use of commodity IT software in control systems made them vulnerable to such attacks [3]. A more recent example is the cyber attack on the Ukrainian power system in the year 2015 which caused a power outage which affected about $200,000$ customers for several

*Reprinted with permission from "Dynamic Watermarking: Active Defense of Networked Cyber-Physical Systems," by B. Satchidanandan and P. R. Kumar in *Proceedings of the IEEE,* vol. 105, no. 2, pp. 219-240, Feb. 2017 by IEEE.

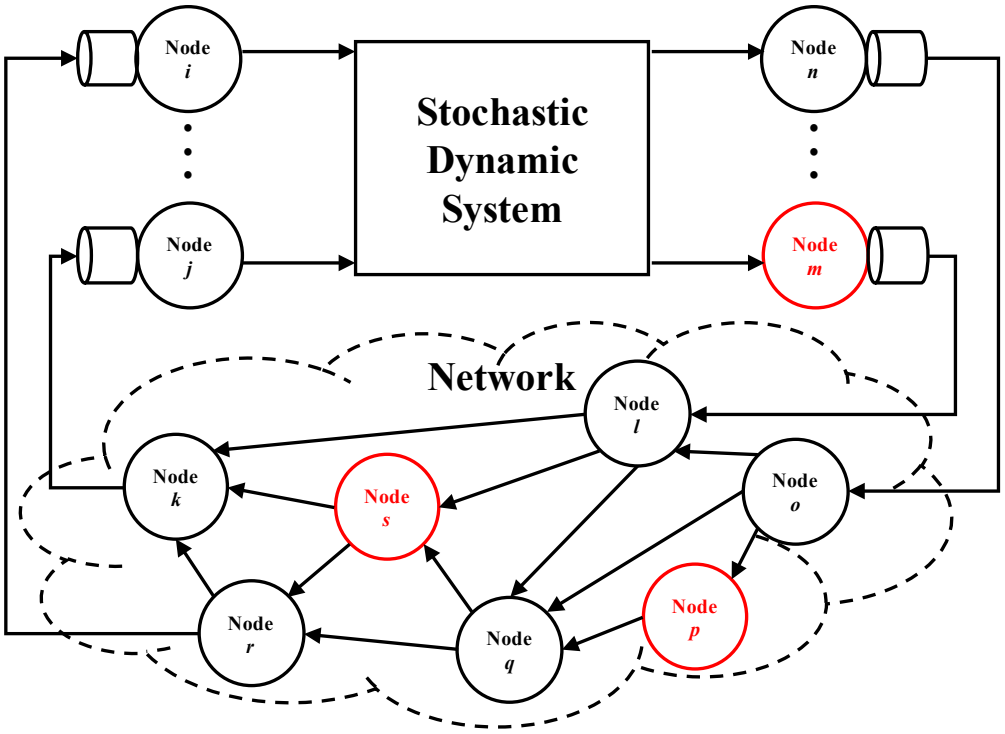

Figure 1.1: A Cyber-Physical System

hours [6]. Another pertinent example is the Stuxnet worm which, in the year 2010, exploited a vulnerability in Microsoft Windows to subvert critical computers controlling centrifuges in Iran's uranium enrichment facility [7]. Having subverted the computers, it issued control commands that caused the centrifuges to operate at abnormally high speeds, causing them to tear themselves apart. In order to keep the attacks undetected by software-implemented alarm routines and officials in the control room, Stuxnet recorded the sensor values in the facility for twenty-one seconds before carrying out each attack, and replayed those twenty-one seconds in a constant loop during the attack. Stuxnet has been claimed to be the first known digital weapon [7], and, since then, cyberwarfare has emerged as a serious concern for cyber-physical systems due to the many advantages it offers to the attacker such as allowing it to remain anonymous, and attack without geographical constraints, etc. Today, the resources required to carry out such attacks on critical infrastructures are generally available [8], underlining the urgent need for the research community to pay attention to the problem.

In this thesis, we examine the problem of detecting attacks on networked cyber-physical systems. These systems can be thought of as having two layers - a physical layer, which consists of the plant, actuators, controllers, and sensors, that interact with physical signals, and a cyber layer, that networks the components of the physical layer. Components in both the cyber layer as well as the physical layer could be subverted or “malicious," as illustrated in Fig. 1.1. While securing the cyber layer is certainly of importance, by itself, it does not address the security of the cyber-physical system as a whole. The Maroochy-Shire incident is a classic illustration of this point, where the malfunctioning of the plant was not the result of an attack on the network layer, but of authorized individuals attacking the physical layer by the issue of improper control commands.

At first glance, it appears as though securing the physical layer is more difficult than securing the cyber layer. In the problem of network security, there is a clear distinction between an honest party and an adversary. Attributes such as credentials and cryptographic keys distinguish honest parties from adversaries. However, when it comes to securing the physical layer, no such demarcation exists. Any authorized party is also a potential adversary.

However, we show in this thesis that what works in favor of securing the physical layer, and what we exploit, is the fact that the actions of every node interfacing with the physical layer get transformed into physical signals, and these signals can be subjected to scrutiny for semantic consistency. To elaborate, consider a physical system consisting of a plant, a set of actuators, and a set of sensors. If the input-output relation of the physical plant is known, then, by observing the signals at various points in the system, it is possible to make inferences about the actions of the sensors and the actuators in the system. By combining this inference with the knowledge of how the various nodes are expected to behave, malicious behavior can potentially be revealed. This intuition will be made precise in subsequent chapters.

The ideas of controllable and unobservable subspaces of a linear dynamical system, introduced by Kalman in [9], play a central role in the theory of control systems. They provide, for example, necessary and sufficient conditions for the existence of a stabilizing control law for any linear dynamical system of interest. Analogous to the notions of controllable and unobservable subspaces,

we begin by examining in this thesis the notions of *"securable"* and *"unsecurable"* subspaces of a linear dynamical system, which we show have operational significance in the context of secure control.

Consider a multiple-input, multiple-output, discrete-time linear dynamical system, an arbitrary subset of whose sensors and actuators may be malicious. The malicious sensors may not truthfully report the measurements that they observe, and the malicious actuators may not apply their control inputs in accordance with the specified control law. Moreover, the malicious actuators and sensors may act in collusion when applying malicious inputs and reporting inaccurate observations. They may strategically tailor them to disrupt the system. In such a setting, even if the system is controllable and observable, the desired control objective of the honest sensors and actuators may not be achievable. The honest nodes in the system may believe the state trajectory to be a certain sequence $\{\mathbf{x}[0], \mathbf{x}[1], \ldots\}$, whereas the actual state trajectory of the system may be very different. It is against this backdrop that we define the notions of securable and unsecurable subspaces of a linear dynamical system as, roughly, the set of states that the system could actually be in, or ever reach, without the honest sensors ever being able to detect, based on their measurements, that the system had visited that state, or that there was any malicious activity in the system. Our results characterize the securable and unsecurable subspaces of a linear system. The unsecurable subspace is an *invariant subspace* in that if the system is initialized in such a subspace, the malicious nodes can synthesize control actions that keep the system in that subspace without the honest nodes ever being able to detect any malicious activity in the system. They may be regarded as the analogs of the controllable and unobservable subspaces reexamined in an era where there is intense interest in cybersecurity of control systems. The aforementioned characterization is amenable to computation. We also address the case of systems with noise, i.e., linear stochastic dynamical systems. We show that the securable and unsecurable subspaces defined in the context of deterministic systems also have operational meaning in the context of stochastic systems.

One way to view these results is as negative or impossibility results which state that given a linear control system with certain malicious sensors and actuators, it is impossible for the honest

sensors in the system to distinguish certain state trajectories from others. Consequently, it may be impossible to guarantee that the system does not reach certain states that are considered "unsafe." An alternate viewpoint is to look at these results from a system designer's perspective. The results could be regarded as providing guidelines to an engineer for designing secure control systems. For example, for a specified amount of resilience required of the control system, typically quantified by a set of Byzantine nodes that the system should tolerate, or for a specification that the system should not visit certain "unsafe" states, the results can be translated into conditions that the securable and unsecurable subspaces should satisfy in order to meet the security specifications. This can potentially constitute a principled approach to design systems that are secure by construction, as opposed to designing systems to maximize a performance metric, and only subsequently installing ad-hoc security measures as an afterthought.

In the next part of this thesis, we propose a technique of *active* defense, termed "Dynamic Watermarking," which enables the designer to expand the securable subspace of a linear system. The fundamental idea of Dynamic Watermarking is this: If an actuator injects into the system a random probing signal that is not disclosed to other nodes in the system, then, combined with the knowledge of the plant dynamics, the actuator expects the signals to appear in transformed ways at various points in the system. Based on the information that the actuator receives from the sensors about the signals at various points, it can potentially infer if there is malicious activity in the system or not.

We develop the technique of watermarking to secure the physical layer of a noisy dynamical system. We examine the protocol whereby honest actuator nodes deliberately superimpose certain stochastically independent probing signals on top of the control law they are intended to apply. We propose specific "tests" that the actuator nodes perform to infer malicious activity, and establish their effectiveness. As a particular illustration of the results, we show that even if all the sensors are malicious, and the actuators have no measurements that they can directly make and have to completely rely on the malicious sensors for *all* their purported measurements, then even in such an adverse environment, the actuators can under appropriate conditions ensure that the additional

distortion on performance that the malicious sensors can cause is of mean-square zero, if they are to remain undetected. Using this approach of active defense, we establish that under appropriate conditions, no matter in what way the adversarial sensors collude, the amount of distortion that they can add without exposing their presence can have an average power of only zero.

As alluded to above, the method we examine is a dynamic version of "watermarking," [10] where certain indelible patterns are imprinted into a medium that can detect tampering [10, 11, 12]. It shows how one can watermark dynamic signals so that one can detect malicious misbehavior on the part of sensors or actuators. On top of a secure communication system, it provides overall security to a cyber-physical system against malicious sensors and actuators.

The rest of this thesis is organized as follows. Chapter 2 presents an overview of related work on this subject. In Chapter 3, we introduce the notions of the securable and unsecurable subspaces of a linear dynamical system, and present a sequence of results that culminate in a characterization these subspaces. These results are the discrete-time analogs of those developed in [13, 14, 15] for continuous-time linear dynamical systems, and some of them are also contained in [16]. Chapter 4 extends the results of Chapter 3 to the case of a perfectly observed stochastic discrete-time linear dynamical system, and shows that the notions of securable and unsecurable subspaces also have operational meaning in this context. Chapter 5 further extends the above results to one of the most general system models used in modern control theory - partially observed stochastic linear dynamical systems with both process and measurement noise. In Chapter 6, we move away from passive defense techniques and introduce the concept of active defense of networked cyberphysical systems, and describe the approach of Dynamic Watermarking. We establish the security guarantees that it provides in a variety of popular control system contexts. Chapter 7 addresses the problem of designing security-guaranteeing dynamic watermarks. Specifically, given a finite-dimensional perfectly-observed linear system with arbitrarily distributed process noise, we derive the necessary and sufficient conditions that the watermark should satisfy in order to provide a fundamental security guarantee. Chapter 8 addresses a scenario in which there are malicious sensors and actuators mixed with honest sensors and actuators, and wherein the honest actuators watermark their

control inputs and conduct certain tests to detect malicious sensors. The state space of any such system can be decomposed into two orthogonal subspaces, called the watermark-securable and the watermark-unsecurable subspaces, such that the malicious sensors and actuators cannot degrade the state estimation performance of the honest sensors and actuators along the watermark-securable subspace. A certain subset of the watermark-securable subspace is characterized in this chapter. Chapter 9 extends the above results to a class of nonlinear systems describing a vehicular cyber-physical system, and presents the results of a laboratory demonstration of Dynamic Watermarking to secure an automated transportation system. Chapter 10 concludes the thesis.

2. RELATED WORK*

The vulnerability and the need to secure critical infrastructure from cyberattacks has been recognized at least as early as 1997 [17]. Subsequent reports [18, 19, 20, 21, 22, 23] cited demonstrated attacks, identifying potential threats, and analyzing the effects of successful attacks on specific systems. The large-scale replacement of proprietary control software and protocols by commodity IT software and protocols by the industry in order to allow for interoperability and rapid scalability has increased the vulnerability of Industrial Control Systems (ICS) to cyberattacks, and roadmaps were prepared to address security of control systems in various sectors such as the energy sector [24], water sector [25], chemical sector [26], and the transportation sector [27].

Some of the initial work [28] has addressed the definition of what constitutes a secure control system. Certain key operational goals such as closed-loop stability and performance metric of interest are noted in [28], and it is proposed that a secure control system must achieve these operational goals even when under attack, or at least cause only a gradual degradation. It also identifies how the problem of secure control of networked systems departs from the traditional problems of network security and information security. In the former, authorized users or insiders can launch attacks on the system causing physical damage, as in the Maroochy-Shire incident. Hence, network and information security measures such as intrusion prevention and detection, authentication, access control, etc., fundamentally cannot address these attacks. Therefore, securing the network does not amount to securing the NCS. In this thesis, we build a framework on top of a secure communication network to secure the NCS. There has been recent work showing how one can indeed build such a communication network that provides provable guarantees on security, throughput as well as delays [29].

*Parts of this chapter are reprinted with permission from “Dynamic Watermarking: Active Defense of Networked Cyber-Physical Systems," by B. Satchidanandan and P. R. Kumar in *Proceedings of the IEEE,* vol. 105, no. 2, pp. 219-240, Feb. 2017 by IEEE.

A theoretical study of secure control benefits from having a model for the adversary, and [30] defines certain adversary and attack models. A popular attack on communication networks is the Denial-of-Service (DoS) attack, in which the adversary floods the communication network with useless packets, rendering it incapable of transporting useful information. Another attack is the deception attack, where an adversary impersonates another node and transmits false information on its behalf. In the context of a networked control system, the adversary could employ a Denial-of-Service (DoS) attack to prevent the controller and actuator from receiving the data required for their operation, and could employ deception to cause an actuator node to issue incorrect actuation signals. A framework to study the evolution of the physical process under DoS and deception attacks is presented in [30].

Classical estimation algorithms such as the Kalman filter can be extended to include the presence of an unreliable network, which may stem from either the characteristics of the network or from adversarial presence, with some packets dropped. This affects the performance of the estimation and control algorithm. Motivated by this scenario, a large body of research has been devoted to addressing the stability of estimation and control algorithms for systems with intermittent observations, c.f. [31, 32, 33, 34, 35] and references therein. Though this line of work does not explicitly model adversarial behavior, some of these ideas have been employed in the literature of secure control. For instance, [36] studies, using some of the machinery developed in [31] in the context of control over lossy communication networks, the effect of a DoS attack on the control performance. Also addressed in the literature is the effect of deception attacks on estimation and control performance. In [37], the effect of false data fed by the compromised sensors to the state estimator is studied. The goal is to characterize the set of all estimation biases that an adversary can inject without being identified. A related work is [38], where fundamental trade-offs between the detection probability (for a fixed false alarm rate) and the estimation error that the adversary can cause, are studied for single-input-single-output systems.

Attack strategies that are designed to be undetected are referred to as stealthy attacks. Reference [16] addresses the problem of taking pro-active measures to reveal stealthy attacks, and also

develops methods to design a system in a manner that renders such attacks infeasible. Characterization of attacks are feasible without detection or identification in continuous-time linear dynamical systems are presented in [13, 14, 15].

Several techniques have also been developed to counter attacks on cyber-physical systems. A technique for correct recovery of the state estimate in the presence of malicious sensors is presented in [39],along with a characterization of the number of malicious sensors that can be tolerated by the algorithm. A well-known attack on control systems is the replay attack, where the adversary records the sensor measurements for a fixed period of time and replays them during the attack so as to maintain the illusion of a normal operating condition. It was shown in [40] that only systems for which the matrix $(A + BL)(I - KC)$ is stable are susceptible to replay attacks, where L is the feedback gain and K is the steady-state Kalman gain. Consequently, a method to secure the system from replay attack is presented in [40, 41, 42, 43]. The fundamental idea is to inject into the actuation signal a component that is not known in advance. Specifically, [40, 42] consider the replay attack, employed in Stuxnet, and introduce a technique, termed Physical Watermarking, wherein the controller commands the actuators to inject into the system a component that is random and not known in advance in order to secure the system against such an attack. To the best of our knowledge, this is the first use of the idea of watermarking. It is shown that by employing Physical Watermarking, the covariance of the innovations process when the system is "healthy" and that when it is under attack, are significantly different, enabling the estimator to detect the attack using a χ_2 detector. This technique is extended in [43] to detect an adversary employing more intelligent attack strategies. Specifically, the adversary is assumed to possess a set of capabilities, based on which a specific attack strategy, consisting of the adversary generating false measurement values that are reported to the estimator, is identified. It is shown that Physical Watermarking can counter such an adversary. Including a random component in the actuation signal would clearly affect the running cost, and [44] develops an optimal policy to switch between cost-centric and security-centric controllers, by formulating the problem as a stochastic game between the system and the adversary. A method to detect false-data injection is presented in [45], where the focus is

on zero-dynamics attacks, attacks which cannot be detected based on input and output measurements. A method to verify the measurements received from multiple sensors is presented in [46], which exploits correlations between sensor measurements and other features to weed out the measurements from malicious sensors. Though not presented in the context of a dynamical system, the ideas presented in [46] could in principle be extended to incorporate system dynamics.

3. THE SECURABLE AND UNSECURABLE SUBSPACES OF A LINEAR CONTROL SYSTEM*

3.1 Introduction

The ideas of controllable and unobservable subspaces of a linear dynamical system, introduced by Kalman, play a central role in the theory of control systems. They provide, for example, necessary and sufficient conditions for the existence of a stabilizing control law for a linear dynamical system of interest. Analogous to the notions of controllable and unobservable subspaces, we examine in this chapter the notions of "securable" and "unsecurable" subspaces of a deterministic linear dynamical system, and show that they have important operational meanings in the context of secure control.

3.2 Problem Formulation

Consider a p^{th} order discrete-time linear dynamical system with m inputs and n outputs described by

$$\begin{aligned} \bar{\mathbf{x}}[t+1] &= A\bar{\mathbf{x}}[t] + B\bar{\mathbf{u}}[t], \\ \bar{\mathbf{y}}[t+1] &= C\bar{\mathbf{x}}[t+1], \\ \bar{\mathbf{x}}[0] &= \mathbf{x}_0. \end{aligned} \tag{3.1}$$

where $\bar{\mathbf{x}}[t] \in \mathbb{R}^p$ denotes the state of the system at time t, $\bar{\mathbf{u}}[t] \in \mathbb{R}^m$ denotes the input applied to the system at time t, $\bar{\mathbf{y}}[t] \in \mathbb{R}^n$ denotes the output of the system at time t, and A, B, and C are real matrices of appropriate dimensions.

Throughout this thesis, we will denote by $\mathbf{z}[t]$ the values *reported* by the sensors at time t.

*Reprinted with permission from "Control systems under attack: The securable and unsecurable subspaces of a linear stochastic system," by B. Satchidanandan and P. R. Kumar in *Emerging Applications of Control and Systems Theory.* Cham, Switzerland: Springer, 2018, pp. 217-228.

If sensor i, $i \in \{1, 2, ..., n\}$, is honest, then $z_i[t] = \bar{y}_i[t]$ for all t. We assume that an arbitrary, known, possibly history-dependent control policy $g = \{g_1, g_2, ...\}$ is in place, and denote by $\bar{\mathbf{u}}^g[t]$ the control policy-specified input at time t, the policy being applied on the reported sequence $\{\mathbf{z}\}$ and not on $\{\bar{\mathbf{y}}\}$. I.e., $\bar{\mathbf{u}}^g[t] = g_t(\mathbf{z}^t)$, where $\mathbf{z}^t \coloneqq [\mathbf{z}^T[0] \;\; \mathbf{z}^T[1] \;\; \dots \;\; \mathbf{z}^T[t]]^T$. Since $\bar{\mathbf{u}}[t]$ denotes the input applied to the system at time t, if actuator i, $i \in \{1, 2, ..., m\}$, is honest, then $\bar{u}_i[t] = \bar{u}_i^g[t]$ for all t.

We assume the adversarial nodes in the system to be near-omniscient, in the sense that at time $t = 0$, they have perfect knowledge of the initial state $\mathbf{x}_0$ of the system. On the other hand, the honest nodes in the system, at any time t, have access only to the measurements $\mathbf{z}^t$ that are reported until that time. Clearly, this assumption represents a worst-case scenario from the point of view of the honest nodes in the system. Consequently, the results presented in this chapter serve as fundamental bounds that apply regardless of the capabilities of the attacker, and in particular, even for systems where the adversary's knowledge may be more limited.

Note that if all the nodes in the system are honest, and if the pair (A, C) is observable, then the nodes can correctly estimate the initial state $\mathbf{x}_0$ of the system by time $p - 1$. Consequently, they can correctly estimate the state $\bar{\mathbf{x}}[t]$ of the system at any time t. However, when there are malicious sensors and/or actuators present in the system, this need not be the case. Specifically, the honest nodes in the system could be under the impression that the state of the system at some time t is $\widehat{\mathbf{x}}[t]$, while in reality, the system could be in state $\bar{\mathbf{x}}[t] \neq \widehat{\mathbf{x}}[t]$. This brings us to the central question that is addressed in this chapter: *Suppose that there are malicious nodes present in the system and that they act in a fashion that keeps them undetected. Suppose also that the honest nodes believe the system's state evolution to be* $\{\widehat{\mathbf{x}}[0], \widehat{\mathbf{x}}[1], \widehat{\mathbf{x}}[2], \dots\}$. *Under these conditions, what are the set of states that the system can actually be in, or ever reach?* This set essentially contains the set of states that the malicious nodes can steer the system to. For this reason, we term this set as the *"unsecurable"* subspace of the system (A, B, C). The orthogonal complement of this is called the *"securable"* subspace. The projection of the uncertain state on this subspace is actually what the honest sensors and actuators believe it is, whether the system is not under attack or is under a

stealthy attack. It is the largest such subspace. A formal definition of securable and unsecurable subspaces is presented in the next section.

3.3 Securable and Unsecurable Subspaces of Linear Control Systems

In order to determine if malicious nodes are present in the system or not, each honest sensor i subjects the reported measurement sequence $\{\mathbf{z}\}$ to the following test. If and only if the test fails (at any time t) does the sensor declare that malicious nodes are present in the system. It raises an alarm only when it is sure that the system is under attack. When there is no evidence of an attack, it does not raise an alarm.

The rest of the chapter follows the notation specified in the appendix of this chapter.

Test: At each time t, check if the reported sequence of measurements up to that time $\mathbf{z}^t$ satisfies the following condition: $\exists \widehat{\mathbf{x}}_0 \in \mathbb{R}^p$ such that,

$$\mathbf{z}^t - F[t-1]\bar{\mathbf{u}}^{g^{t-1}} = \Gamma[t]\widehat{\mathbf{x}}_0. \tag{3.2}$$

Proposition 1. *If all the nodes in the system are honest, the reported measurements $\{\mathbf{z}\}$ pass (3.2) at each time t. Conversely, if the reported measurements $\{\mathbf{z}\}$ pass (3.2) at each time t, then there exists an initial state $\mathbf{x}[0]$ such that $\mathbf{z}[t]$ is the output of the system at time t under control $\{\bar{\mathbf{u}}^g\}$, and so, there is no definitive reason for the honest sensor to declare that malicious nodes are present in the system.*

Proof. First assume that all the nodes in the system are honest. Then, $\mathbf{z}[t] = \bar{\mathbf{y}}[t]$ for all t, and so, $\mathbf{z}^t - F[t-1]\bar{\mathbf{u}}^{g^{t-1}} = \bar{\mathbf{y}}^t - F[t-1]\bar{\mathbf{u}}^{g^{t-1}} = \Gamma[t]\mathbf{x}_0$. Consequently, (3.2) reduces to checking if $\exists \widehat{\mathbf{x}}_0$ such that $\forall t$, $\Gamma[t]\mathbf{x}_0 = \Gamma[t]\widehat{\mathbf{x}}_0$, which is clearly true.

Next assume that the reported measurement sequence $\{\mathbf{z}\}$ satisfies (3.2). Then, if the initial state $\bar{\mathbf{x}}[0] = \widehat{\mathbf{x}}_0$, and $\bar{\mathbf{u}} \equiv \bar{\mathbf{u}}^g$ (i.e., $\bar{\mathbf{d}} \equiv 0$), we would have $\bar{\mathbf{y}}^t = \Gamma[t]\widehat{\mathbf{x}}_0 + F[t-1]\bar{\mathbf{u}}^{g^{t-1}} = \mathbf{z}^t$, where the last equality follows from (3.2). □

In what follows, we assume that the measurements reported by the malicious sensors pass the above test, and examine the limits of what the malicious nodes can do under this constraint.

Since the reported measurements $\{\mathbf{z}\}$ pass (3.2), it follows in particular that $\exists \widehat{\mathbf{x}}_0 \in \mathbb{R}^p$ such that $\forall t$,

$$\mathbf{z}^{t-1} - F[t-2]\bar{\mathbf{u}}^{g^{t-2}} = \Gamma[t-1]\widehat{\mathbf{x}}_0, \tag{3.3}$$

$$\bar{\mathbf{y}}_H[t] - \sum_{i=0}^{t-1} C_H A^i B \bar{\mathbf{u}}^g[t-1-i] = C_H A^t \widehat{\mathbf{x}}_0. \tag{3.4}$$

The following proposition is a (partial) converse of the above statement.

Proposition 2. *Suppose that there exist $\widehat{\mathbf{x}}_0$, $\mathbf{z}_M^{\tau-1}$, and $\bar{\mathbf{d}}^{\tau-1}$ such that (3.3) and (3.4) hold for $t = \tau$. Then, there is a vector $\mathbf{z}_M[\tau]$ that satisfies Test (3.2) at time τ.*

Proof. Consider $\mathbf{z}_M[\tau] = C_M A^\tau \widehat{\mathbf{x}}_0 + \sum_{i=0}^{\tau-1} C_M A^i B \bar{\mathbf{u}}^g[\tau-1-i]$. It is straightforward to verify that it satisfies (3.2). □

The above proposition states that it is sufficient for the malicious nodes to consider strategies that only ensure “consistency" at the outputs of the honest sensors. The outputs to be reported by the malicious sensors can be fabricated accordingly.

The next proposition, along with Theorem 2, shows that one can consider a simpler system consisting of only malicious actuators, honest sensors, and a control policy that is identically zero, and translate the conclusion obtained from the analysis of such a system to the more general system (3.1). In other words, one can dispense with the honest actuators and malicious sensors. There is no loss of generality in assuming that the control policy is identically equal to zero, and that the system has only honest sensors and malicious actuators.

Given the system described by (3.1), consisting of honest and malicious nodes as described before, consider the following reduction of the system where all sensors are honest, all actuators

are malicious, and the control policy is identically equal to zero:

$$\begin{aligned} \mathbf{x}[t+1] &= A\mathbf{x}[t] + B_M\mathbf{d}[t], \\ \mathbf{y}_H[t+1] &= C_H\mathbf{x}[t+1], \\ \mathbf{x}[0] &= \mathbf{x}_0. \end{aligned} \tag{3.5}$$

where $\mathbf{y}_H[t]$ are the measurements observed by the (honest) sensors at time t, $\mathbf{d}[t]$ are the inputs applied by the (malicious) actuators at time t. We will refer to system (3.5) as the "reduced system" of system (3.1), or simply the "reduced system" when there is no ambiguity. Note that the reduced system has the same state space as its parent system (3.1), and is also initialized with the same state as its parent. It is only the inputs and the outputs of the systems that are different. As before, the malicious actuators are assumed to be near-omniscient so that they have perfect knowledge of the initial state $\mathbf{x}_0$. It is also worth noting that the reduced system may be neither controllable nor observable, even if the parent system is. For the reduced system, Test (3.2) reduces to the following, and is performed by the (honest) sensors.

Test for the reduced system: Check if $\exists\widetilde{\mathbf{x}}_0 \in \mathbb{R}^p$ such that for all t,

$$\mathbf{y}_H^t = \Gamma_H[t]\widetilde{\mathbf{x}}_0. \tag{3.6}$$

Proposition 3. *Suppose that there exists a sequence $\{\mathbf{d}\}$ for the reduced system satisfying test (3.6). Then, if the malicious actuators in the parent system (3.1) inject $\{\bar{\mathbf{d}}\} \equiv \{\mathbf{d}\}$, there exist fabricated measurements $\{\mathbf{z}_M\}$ that can be reported by the malicious sensors in the parent system that pass Test (3.2) with $\widehat{\mathbf{x}}_0 = \widetilde{\mathbf{x}}_0$.*

Proof. For the reduced system, we have

$$\mathbf{y}_H[t] = C_H A^t \mathbf{x}_0 + \sum_{i=0}^{t-1} C_H A^i B_M \mathbf{d}[t-1-i]. \tag{3.7}$$

Now, suppose for induction that there exist measurements $\mathbf{z}_M[0], \mathbf{z}_M[1], \ldots, \mathbf{z}_M[t-1]$ that the

malicious sensors can report for system (3.1) when the malicious actuators inject $\bar{\mathbf{d}}[i] = \mathbf{d}[i]$, $i = 0, 2, ..., t-2$, such that the reported measurements pass test (3.2) up to time $t-1$ with $\widehat{\mathbf{x}}_0 = \widetilde{\mathbf{x}}_0$. The base case of $t = 1$ holds since the malicious sensors in the parent system can report $\mathbf{z}_M[0] = C_M\widetilde{\mathbf{x}}_0$. This amounts to assuming that (3.3) holds with $\widehat{\mathbf{x}}_\mathbf{0} = \widetilde{\mathbf{x}}_0$. Now, if the malicious actuators in the parent system inject, at time $t-1$, $\bar{\mathbf{d}}[t-1] = \mathbf{d}[t-1]$, then,

$$\bar{\mathbf{y}}_H[t] = C_H A^t \mathbf{x}_0 + \sum_{i=0}^{t-1} C_H A^i B \bar{\mathbf{u}}^g[t-1-i] + \sum_{i=0}^{t-1} C_H A^i B_M \mathbf{d}[t-1-i].$$

Substituting (3.7) in the above gives

$$\bar{\mathbf{y}}_H[t] = \mathbf{y}_H[t] + \sum_{i=0}^{t-1} C_H A^i \bar{\mathbf{u}}^g[t-1-i].$$

Since the output of the reduced system satisfies (3.6), we have $\mathbf{y}_H[t] = C_H A^t \widetilde{\mathbf{x}}_0$. Substituting this into the above equation gives

$$\bar{\mathbf{y}}_H[t] - \sum_{i=0}^{t-1} C_H A^i \bar{\mathbf{u}}^g[t-1-i] = C_H A^t \widetilde{\mathbf{x}}_0,$$

which satisfies (3.4) for $\widehat{\mathbf{x}}_\mathbf{0} = \widetilde{\mathbf{x}}_0$. The desired result follows from Proposition 2. □

The following definition is of central importance.

Definition 1. *Consider a system (A, B, C) of the form (3.1) with initial state $\mathbf{x}_0$. The* unsecurable subspace for state $\mathbf{s}_0$ *of the system is the maximal set of states $\mathcal{V}(\mathbf{s}_0)$ such that for each $\mathbf{v} \in \mathcal{V}(\mathbf{s}_0)$, there exist $t, \{\bar{\mathbf{d}}\}, \{\mathbf{z}_M\}$ such that $\bar{\mathbf{x}}[t] = \mathbf{v}$ and (3.2) holds for $\widehat{\mathbf{x}}_0 = \mathbf{s}_0$.*

In particular, for the reduced system (A, B_M, C_H), the unsecurable subspace for state $\mathbf{s}_0$ is the maximal set of states $\mathcal{V}_R(\mathbf{s}_0)$ such that for each $\mathbf{v} \in \mathcal{V}_R(\mathbf{s}_0)$, there exist $t, \{\mathbf{d}\}$ such that $\mathbf{x}[t] = \mathbf{v}$ and (3.6) holds for $\widetilde{\mathbf{x}}_0 = \mathbf{s}_0$.

In other words, the unsecurable space for $\mathbf{s}_0$ is the set of states that the system can be in if the honest nodes are deceived into inferring the initial state as $\mathbf{s}_0$. Note that for the reduced system (or

even the parent system for that matter), the vector sequence that the honest nodes believe is the state trajectory is completely determined by what they believe the initial state is (or the set that they believe the initial state lies in, if the system is unobservable). If the unsecurable subspace is of dimension greater than zero, it (i) states that the malicious nodes cannot distort certain linear combinations of the state without being detected, and (ii) specifies those linear combinations that are "intact."

The following theorem characterizes the unsecurable subspace and provides an algorithm to compute it.

Theorem 1. *Consider a reduced system (A, B_M, C_H) of the form (3.5). For such a system,*

(i) The unsecurable subspace $\mathcal{V}_R(0)$ for state 0 is the maximal set $W \subseteq \mathbb{R}^p$ such that $\forall \mathbf{w} \in W$,

(a) $C_H\mathbf{w} = 0$, and

(b) $\exists \mathbf{d}$ such that $A\mathbf{w} + B_M\mathbf{d} \in W$.

(ii) The unsecurable subspace for state $\mathbf{s}_0$, $\mathcal{V}_R(\mathbf{s}_0)$, is

$$\mathcal{V}_R(\mathbf{s}_0) = \{\mathbf{s}_0 + \mathbf{w} : \mathbf{w} \in \mathcal{V}_R(0)\}. \tag{3.8}$$

Proof. We begin with the following lemma.

Lemma 1. *The set W is a subspace.*

Proof. Let $\mathbf{w}_1, \mathbf{w}_2 \in W$. We have from the definition of W, that $H_M[t-1]\mathbf{d}_i^{t-1} = \Gamma_H[t]\mathbf{w}_i$ has a solution $\mathbf{d}_i^{t-1}$ for all t, $i = 1, 2$. It follows that the equation $H_M[t-1]\mathbf{d}^{t-1} = \Gamma_H[t](k_1\mathbf{w}_i + k_2\mathbf{w}_2)$ has a solution $\mathbf{d}^{t-1}$, viz., $\mathbf{d}^{t-1} = k_1\mathbf{d}_1^{t-1} + k_2\mathbf{d}_2^{t-1}$ for all t, which in turn implies that $k_1\mathbf{w}_1 + k_2\mathbf{w}_2 \in W$. □

We now show that W is equal to $\mathcal{V}_R(0)$. The crux of the argument is that W is an invariant subspace in the following sense.

Lemma 2. *If the system's state visits W at any time t, then the malicious actuators can synthesize control actions that keep the state in W at all subsequent times.*

Proof. We show this via induction. Let $\mathbf{w} \in W$, and let $\mathbf{x}[t] = \mathbf{w}$, which also serves as the base case for induction. Assume for induction that $\mathbf{x}[\tau] \in W$, where $\tau \geq t$ is a fixed time. Then, $\mathbf{x}[\tau+1] = A\mathbf{x}[\tau] + B_M\mathbf{d}[\tau]$. Since $\mathbf{x}[\tau] \in W$, it follows from the definition of W that there exists a control choice $\mathbf{d}$ for $\mathbf{d}[\tau]$ such that $A\mathbf{x}[\tau] + B_M\mathbf{d}[\tau] \in W$, implying that $\mathbf{x}[\tau+1] \in W$. □

Remark: Owing to the above Lemma, W is called the "controlled invariant subspace" in linear system theory [47].

Now, suppose that $\mathbf{x}[0] = \mathbf{w}$ and $\mathbf{w} \in W$. We then have from Lemma 2 that there exists a sequence $\{\mathbf{d}\}$ that the malicious actuators can apply as inputs such that $\mathbf{x}[t] \in W$ for all t. Since $W \subseteq \mathrm{N}(C_H)$ by definition, we have $\mathbf{y}_H[t] = C_H\mathbf{x}[t] = 0$ for all t. Consequently, (3.6) holds for $\widetilde{\mathbf{x}}_0 = 0$, and it follows from Definition 1 that $\mathbf{w} \in \mathcal{V}_R(0)$. Hence, $W \subseteq \mathcal{V}_R(0)$.

Now suppose that $\mathbf{v} \in \mathcal{V}_R(0)$. We then have from Definition 1 that $\exists\{\mathbf{d}\}$ that the malicious actuators can apply as inputs to the system such that $\mathbf{x}[t] = \mathbf{v}$ for some t and (3.6) holds for $\widetilde{\mathbf{x}}_0 = 0$. This implies that $\mathbf{y}_H[t] = 0$ for all t. Since $0 = \mathbf{y}_H[t] = C_H\mathbf{x}[t] = C_H\mathbf{v}$, we have that $\mathbf{v} \in \mathrm{N}(C_H)$, satisfying the first condition to be an element of W. Since $\{\mathbf{d}[t], \mathbf{d}[t+1], ...\}$ is a sequence that the malicious actuators can apply such that $\mathbf{x}[t'] \in \mathrm{N}(C_H)$ for all $t' \geq t$, $\mathbf{v}$ satisfies the second condition to be an element of W. Therefore, $\mathbf{v} \in W$, and $\mathcal{V}_R(0) \subseteq W$. Combining the two results, we have $W = \mathcal{V}_R(0)$.

(ii) Let $\mathbf{v} \in \{\mathbf{s}_0 + \mathbf{w} : \mathbf{w} \in \mathcal{V}_R(0)\}$, and let $\mathbf{x}[0] = \mathbf{v}$. Then, $\mathbf{y}_H^t = \Gamma_H[t]\mathbf{v} + H_M[t-1]\mathbf{d}^{t-1} = \Gamma_H[t]\mathbf{s}_0 + \Gamma_H[t]\mathbf{w} + H_M[t-1]\mathbf{d}^{t-1}$ for all t. Since $\mathbf{w} \in \mathcal{V}_R(0)$, it follows from the definition of $\mathcal{V}_R(0)$ that there exists sequence $\{\mathbf{d}\}$ so that $\Gamma_H[t]\mathbf{w} + H_M[t-1]\mathbf{d}^{t-1} = 0$ for all t. Therefore, if the actuators inject such a sequence $\{\mathbf{d}\}$, then, $\mathbf{y}_H^t$ reduces to $\mathbf{y}_H^t = \Gamma_H[t]\mathbf{s}_0$. Therefore, (3.6) holds with $\widetilde{\mathbf{x}}_0 = \mathbf{s}_0$, and so, $\{\mathbf{s}_0 + \mathbf{w} : \mathbf{w} \in \mathcal{V}(0)\} \subseteq \mathcal{V}_R(s_0)$.

Next, let $\mathbf{v} \in \mathcal{V}_R(s_0)$. Then, we have from the definition of $\mathcal{V}_R(\mathbf{s}_0)$ that $\exists\{\mathbf{d}'\}, \tau$ such that $\mathbf{x}[\tau] = \mathbf{v}$ and $\Gamma_H[t]\mathbf{s}_0 = \mathbf{y}_H^t$ for all t. This in turn implies that $\exists\{\mathbf{d}\}$ such that $\mathbf{x}_0 = \mathbf{v}$ and

$\Gamma_H[t]\mathbf{s}_0 = \mathbf{y}_H^t$ for all t. Also, when $\mathbf{x}_0 = \mathbf{v}$, we have for all t, $\mathbf{y}_H^t = \Gamma_H[t]\mathbf{v} + H_M[t-1]\mathbf{d}^{t-1}$. Combining the two, we have that there exists $\{\mathbf{d}\}$ such that $\Gamma_H[t]\mathbf{s}_0 = \Gamma_H[t]\mathbf{v} + H_M[t-1]\mathbf{d}^{t-1}$ for all t. This means that $\mathbf{v}$ solves, for all t,

$$[\Gamma_H[t] \;\; H_M[t-1]] \begin{bmatrix} \mathbf{v} \\ \mathbf{d}^{t-1} \end{bmatrix} = [\Gamma_H[t] \;\; H_M[t-1]] \begin{bmatrix} \mathbf{s}_0 \\ 0 \end{bmatrix},$$

so that for all t,

$$\begin{bmatrix} \mathbf{v} \\ \mathbf{d}^{t-1} \end{bmatrix} = \begin{bmatrix} \mathbf{s}_0 \\ 0 \end{bmatrix} + \widetilde{\mathbf{w}},$$

where $\widetilde{\mathbf{w}} \in \mathrm{N}([\Gamma_H[t] \;\; H_M[t-1]])$. Denote by $\mathbf{w}$ the first p entries of $\widetilde{\mathbf{w}}$, and it follows from the definition of $\mathcal{V}_R(0)$ that $\widetilde{\mathbf{w}} \in \mathcal{V}_R(0)$. Hence, $\mathbf{v}$ must be of the form $\mathbf{s}_0 + \mathbf{w}$, $\mathbf{w} \in \mathcal{V}_R(0)$, and hence, $\mathcal{V}_R(s_0) \subseteq \{\mathbf{s}_0 + \mathbf{w} : \mathbf{w} \in \mathcal{V}(0)\}$.

Combining the two results, we have $\mathcal{V}_R(\mathbf{s}_0) = \{\mathbf{s}_0 + \mathbf{w} : \mathbf{w} \in \mathcal{V}(0)\}$. □

The following theorem translates the above conclusions obtained from the reduced system (3.5) to the original system (3.1) that is of interest.

Theorem 2. *The unsecurable subspace $\mathcal{V}(\mathbf{s}_0)$ for the system (A, B, C) is the same as the unsecurable subspace $\mathcal{V}_R(\mathbf{s}_0)$ for its reduction (A, B_M, C_H).*

Proof. Let $\mathbf{v} \in \mathcal{V}_R(\mathbf{s}_0)$. Then, it follows form the definition of $\mathcal{V}_R(\mathbf{s}_0)$ that for the reduced system (3.5), there exists $\{\mathbf{d}\}$ that can be applied by the actuators so that (3.6) is satisfied for $\widetilde{\mathbf{x}}_0 = \mathbf{s}_0$ when $\mathbf{x}_0 = \mathbf{v}$. Therefore, by Proposition 3, $\mathbf{v} \in \mathcal{V}(\mathbf{s}_0)$, and so, $\mathcal{V}_R(\mathbf{s}_0) = \mathcal{V}(\mathbf{s}_0)$.

Next, let $\mathbf{v} \in \mathcal{V}(\mathbf{s}_0)$. Then, from the definition of $\mathcal{V}(\mathbf{s}_0)$, we have for system (3.1) that $\exists\{\mathbf{d}\}, \{\mathbf{z}_M\}$ such that for all t, $\mathbf{z}^t = \Gamma[t]\mathbf{s}_0 + F[t-1]\bar{\mathbf{u}}^{g^{t-1}}$ when $\mathbf{x}_0 = \mathbf{v}$. This implies that $\bar{\mathbf{y}}_H^t = \Gamma_H[t]\mathbf{s}_0 + H[t-1]\bar{\mathbf{u}}^{g^{t-1}}$. Since we also have $\bar{\mathbf{y}}_H^t = \Gamma_H[t]\mathbf{v} + H[t-1]\bar{\mathbf{u}}^{g^{t-1}} + H_M[t-1]\mathbf{d}^{t-1}$, substituting

this in the previous equation gives

$$\Gamma_H[t]\mathbf{v} + H_M[t-1]\mathbf{d}^{t-1} = \Gamma_H[t]\mathbf{s}_0. \tag{3.9}$$

Now, if the actuators apply the above sequence $\{\mathbf{d}\}$ to the reduced system (with initial state $\mathbf{v}$), we have for each t, $\mathbf{y}_H^t = \Gamma_H[t]\mathbf{v} + H_M[t-1]\mathbf{d}^{t-1} = \Gamma_H[t]\mathbf{s}_0$, where the last equality follows from the (3.9). Hence, (3.6) is satisfied with $\widetilde{\mathbf{x}}_0 = \mathbf{s}_0$, and hence, $\mathcal{V}(\mathbf{s}_0) \subseteq \mathcal{V}_R(\mathbf{s}_0)$.

Combining the two results, we have $\mathcal{V}(\mathbf{s}_0) = \mathcal{V}_R(\mathbf{s}_0)$. □

The characterization of $\mathcal{V}_R(0)$ given in Theorem 1 allows one to use standard algorithms that compute $(A, \mathcal{R}(B_M))-$controlled invariant subspaces of a linear dynamical system for computing its unsecurable subspace.

Definition 2. *The* securable subspace S *of a discrete-time linear dynamical system of the form (3.1) is the orthogonal complement of* $\mathcal{V}(0)$*, the unsecurable subspace of the zero state.*

The securable subspace has the interpretation of the maximal set of states that the malicious nodes cannot steer the system to without leaving a nonzero trace at the output of the honest sensors. Chapters 4 and 5 examine the performance of a stochastic linear dynamical system in the securable subspace, which provide further operational meaning to it.

3.4 Concluding Remarks

In this chapter, we have introduced the notions of the securable and the unsecurable subspaces of a linear dynamical system. The unsecurable subspace has the interpretation as a set of states that the system could actually be in, or ever reach, as a consequence of malicious actions of the adversarial nodes, without the honest sensors in the system ever detecting definitively any malicious activity. This is an invariant subspace in the sense that once the state of the system enters this space, the malicious sensor and actuator nodes in the system can collude to keep the system in this space forever without the honest sensors ever being able to confirm any malicious activity based on their own observations or the ones being reported to them. The orthogonal complement of this

subspace, the securable subspace, has the interpretation in the context of deterministic systems as the set of states that the malicious nodes cannot steer the system to without leaving a nonzero trace at the output of the honest sensors. A characterization of these subspaces, and an algorithm to compute them for any linear system and any combination of malicious sensors and actuators is obtained by a standard recourse to geometric control methods. These subspaces also have relevance to the case where the system is noisy, and these are elaborated in the following chapters.

Notation

The following notation is used throughout this chapter.

1. Let $s_1 < s_2 < ... < s_{h_s}$ denote the indices of the honest sensors, $\psi_1, \psi_2, ..., \psi_{m_s}$ denote those of the malicious sensors, and $a_1 < a_2 < ... < a_{m_a}$ denote those of the malicious actuators. Then,

   - C_H is the $h_s \times p$ matrix whose i^{th} row is the s_i^{th} row of C, $i = 1, 2, ..., h_s$,
   - B_M is the $p \times m_a$ matrix whose i^{th} column is the a_i^{th} column of B, $i = 1, 2, ..., m_a$,
   - $\bar{\mathbf{y}}_H[t]$ is the $h_s \times 1$ vector whose i^{th} component is the s_i^{th} entry of $\bar{\mathbf{y}}[t]$, $i = 1, 2, ..., h_s$,
   - $\mathbf{z}_M[t]$ is the $m_s \times 1$ vector whose i^{th} entry is the ψ_i^{th} entry of $\mathbf{z}[t]$, $i = 1, 2, ..., m_s$,
   - $\bar{\mathbf{d}}[t]$ is the $m_a \times 1$ vector whose i^{th} component is $\bar{d}_i[t] \coloneqq \bar{u}_{a_i}[t] - \bar{u}^g_{a_i}[t]$, $i = 1, 2, ..., m_a$.

2. $\mathbf{x}^t$ denotes $[\mathbf{x}^T[0] \;\; \mathbf{x}^T[1] \;\; \ldots \;\; \mathbf{x}^T[t]]^T$.

3. $\Gamma[t] \coloneqq [C^T \;\; (CA)^T \;\; (CA^2)^T \;\; \ldots \;\; (CA^t)^T]^T$.

4. $\Gamma_H[t] \coloneqq [(C_H)^T \;\; (C_H A)^T \;\; \ldots \;\; (C_H A^t)^T]^T$.

5.

$$F[t] \coloneqq \begin{bmatrix} 0 & 0 & \dots & 0 \\ CB & 0 & \dots & 0 \\ CAB & CB & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ CA^tB & CA^{t-1}B & \dots & CB \end{bmatrix},$$

6.

$$H[t] \coloneqq \begin{bmatrix} 0 & 0 & \dots & 0 \\ C_HB & 0 & \dots & 0 \\ C_HAB & C_HB & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ C_HA^tB & C_HA^{t-1}B & \dots & C_HB \end{bmatrix},$$

$$H_M[t] \coloneqq \begin{bmatrix} 0 & 0 & \dots & 0 \\ C_HB_M & 0 & \dots & 0 \\ C_HAB_M & C_HB_M & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ C_HA^tB_M & C_HA^{t-1}B_M & \dots & C_HB_M \end{bmatrix}.$$

7. $\mathrm{N}(\cdot)$ denotes the null space of a matrix, and $\mathcal{R}(\cdot)$ denotes the range space of a matrix.

4. THE SECURABLE SUBSPACE OF A LINEAR STOCHASTIC SYSTEM WITH MALICIOUS SENSORS AND ACTUATORS*

4.1 Introduction

In the previous chapter, we introduced the notions of securable and unsecurable subspaces of a deterministic linear dynamical system. In this chapter, we view these subspaces from the standpoint of stochastic linear dynamical systems, and establish important operational meanings for them in the stochastic context. Of the many potential applications that the results developed in this chapter may have, perhaps the most significant is in synthesizing control systems that are provably secure by design.

The rest of the chapter is organized as follows. Section 4.2 formulates the problem, and recapitulates the notions of securable and unsecurable subspaces of a linear dynamical system. Section 4.3 is devoted to establishing certain key geometric properties of the securable subspace. These properties are exploited in Section 4.4 to establish an operational meaning for the securable and the unsecurable subspace in the context of fully-observed stochastic systems.

Notation: Given a vector $\mathbf{x}$ and a subspace W, x_i denotes the i^{th} entry of $\mathbf{x}$ and $\mathbf{x}_W$ denotes the projection of $\mathbf{x}$ on the subspace W. Given a matrix A, $A_{i,\times}$ denotes the i^{th} row of A, and $A_{\times,i}$ denotes its i^{th} column. Given a vector $\mathbf{x}$, we denote by $\mathbf{x}^{TT}$ the expression $\mathbf{x}^T\mathbf{x}$, and by $\mathbf{x}_W^{TT}$, the expression $\mathbf{x}_W^T\mathbf{x}_W$.

*Reprinted with permission from "The securable subspace of a linear stochastic system with malicious sensors and actuators," by B. Satchidanandan and P. R. Kumar in *Proceedings of the 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton),* pp. 911-917, 2017, IEEE.

4.2 Securable and Unsecurable Subspaces of Linear Dynamical Systems

Consider an $m \times n$ stochastic linear dynamical system of order p described by

$$\begin{aligned} \mathbf{x}[t+1] &= A\mathbf{x}[t] + B\mathbf{u}[t] + \mathbf{w}[t+1], \\ \mathbf{y}[t+1] &= C\mathbf{x}[t+1], \end{aligned} \tag{4.1}$$

where $\mathbf{x}[t] \in \mathbb{R}^p$ is the state of the system, $\mathbf{u}[t] \in \mathbb{R}^m$ is the input applied to the system, $\mathbf{y}[t] \in \mathbb{R}^n$ is the output of the system, $\mathbf{w}[t] \sim \mathcal{N}(0, \sigma_w^2 I)$ is i.i.d across time and denotes the process noise. The matrices A, B, and C are assumed to be known. While the results pertaining to the stochastic case (contained in Section 4.4) assume that $C = I$, the results contained in this section and Section 4.3 hold for arbitrary C, unless otherwise specified.

As mentioned before, throughout this thesis, we denote by $\mathbf{z}[t]$ the measurements reported by the sensors at time t. If sensor $i, i \in \{1, \ldots, n\}$ is honest, then, $z_i[t] \equiv x_i[t]$. A malicious sensor may report $z_i[t] \neq x_i[t]$.

We suppose that a known but arbitrary and possibly history-dependent control policy $\{g_t^i\}, i = 1, \ldots, m$ is in place, so that actuator i, if honest, applies the input $g_t^i(\mathbf{z}^t)$ at time t, where $\mathbf{z}^t \coloneqq (\mathbf{z}[0], \ldots, \mathbf{z}[t])$. Note that since the honest actuators may not have access to the actual output sequence $\mathbf{x}^t$ due to the presence of malicious sensors, they apply the control law on the reported measurements rather than the actual measurements. The malicious actuators, however, may not adhere to the control law, and can apply any input that they please.

Let $a_1, \ldots, a_{m_a}$ denote the indices of the m_a malicious actuators, $a_i \in \{1, \ldots, m\}, i = 1, \ldots, m_a$. Define for malicious actuator a_i, the *input distortion* at time t as $d_{a_i}[t] \coloneqq u_{a_i}[t] - g_t^{a_i}(\mathbf{z}^t)$, which is the difference between the actual input that it applies and the control law-specified input that it is supposed to apply. Then, the system evolves in closed loop as

$$\begin{aligned} \mathbf{x}[t+1] &= A\mathbf{x}[t] + Bg_t(\mathbf{z}^t) + B^M\mathbf{d}[t] + \mathbf{w}[t+1], \\ \mathbf{y}[t+1] &= C\mathbf{x}[t+1], \end{aligned} \tag{4.2}$$

where $\mathbf{d}[t] \coloneqq [d_{a_1}[t] \ \dots \ d_{a_{m_a}}[t]]^T$ is the vector containing the distortions of the malicious actuators at time t, and $B^M \coloneqq [B_{\times,a_1} \ \dots \ B_{\times,a_{m_a}}]$ is the submatrix of B formed by the columns corresponding to the malicious actuators.

Denote by $o_1, \dots, o_{h_s}$ the indices of the h_S honest sensors in the system, $o_i \in \{1, \dots, n\}, i = 1, \dots, h_s$. Define $C^H \coloneqq [C^T_{o_1,\times} \ \dots \ C^T_{o_{h_s},\times}]^T$, which is the submatrix of C formed by the rows corresponding to the honest sensors.

4.2.1 Securable and Unsecurable Subspaces

Definition 3. *For the system (4.1) consisting of some malicious sensors and actuators, the* unsecurable subspace for state 0, *or simply the* unsecurable subspace*, is the maximal subset V of the state space such that $\forall \mathbf{v} \in V$,*

1. $C^H \mathbf{v} = 0$,

2. $\exists \mathbf{d}$, *possibly dependent on* $\mathbf{v}$, *such that* $A\mathbf{v} + B^M \mathbf{d} \in V$.

It is the maximal $(A, B^M)-$invariant subspace contained in $\mathcal{N}(C^H)$.

Definition 4. *The* securable subspace *S of system (4.1) is the orthogonal compliment of its unsecurable subspace. I.e.,*

$$S \coloneqq V^{\perp}.$$

An operational meaning for these subspaces in the context of deterministic linear dynamical systems is presented in [48]. Specifically, in deterministic systems, the unsecurable subspace has the interpretation as the set of states that the malicious nodes can steer the system to ever, without the honest nodes ever detecting that the system had visited that state or that there is any malicious activity in the system. This interpretation does not extend to systems that are stochastic in nature, such as (4.1). In the following sections, we establish an operational meaning for the securable subspace for stochastic linear dynamical systems described by (4.1) for the case of a fully observed system, i.e., $C = I$. Specifically, we show that the securable subspace is the maximal subspace of the state space along which the state estimation error of the honest nodes is of zero covariance,

where the state estimate $\widehat{\mathbf{x}}[t]$ of the honest nodes at time t is simply $\widehat{\mathbf{x}}[t] = \mathbf{z}[t]$, and their state estimation error at time t is $\widehat{\mathbf{x}}[t] - \mathbf{x}[t]$.

To elaborate, note that if all the sensors in a fully observed system were honest, then the state estimation error of the honest sensors and actuators would be identically zero. However, since certain sensors could be malicious, the honest nodes in the system may not have perfect estimates of the state, thereby incurring a state estimation error at each time. However, as we show in the subsequent sections, *this state estimation error, when projected on the securable subspace, has zero covariance*. One of the implications of this is that if the malicious nodes wish to remain undetected, then no matter what attack strategy they employ, they cannot distort the projection of the *actual* state on the securable subspace beyond adding a zero-power sequence.

4.3 The Geometry of the Securable Subspace

This section is devoted to establishing certain key structural properties of the securable subspace that are used in Section 4.4 to establish its operational meaning.

Definition 5. *A* $k-$step unsecurable subspace*, denoted by* $\mathcal{U}_k$*, is defined as*

$$\begin{aligned}\mathcal{U}_k \coloneqq \{\mathbf{v} :\ & \exists \mathbf{d}_0, \ldots, \mathbf{d}_{k-2} \ \ s.t. \\ & C^H \mathbf{v} = 0, \\ & C^H (A\mathbf{v} + B^M \mathbf{d}_0) = 0, \\ & C^H (A^2\mathbf{v} + AB^M \mathbf{d}_0 + B^M \mathbf{d}_1) = 0, \\ & \quad \vdots \\ & C^H (A^{k-1}\mathbf{v} + \ldots + B^M \mathbf{d}_{k-2}) = 0.\}, \end{aligned} \tag{4.3}$$

for $k = 1, \ldots, r$*, where* $r \coloneqq \min\{i \in \mathbb{N} : \mathcal{U}_i = V\}$.

For finite-dimensional linear systems, it can be shown that $r < \infty$ (in fact, $r \leq p$). Note also that $\mathcal{U}_1 = \mathcal{N}(C^H)$.

Intuitively, $\mathcal{U}_k$ is the set of states which, if the system is initialized at, can be made indistin-

guishable to the honest sensors from the zero initial state for *at least* k epochs by some choice of inputs of the malicious actuators.

Note from (4.3) that the $(k+1)-$step unsecurable subspace contains the $k-$step unsecurable subspace, i.e., the subspaces are monotone decreasing in k. The $r-$step unsecurable subspace is simply the unsecurable subspace:

$$V = \mathcal{U}_r \subset \mathcal{U}_{r-1} \subseteq \ldots \subseteq \mathcal{U}_1 \subseteq \mathcal{X}, \tag{4.4}$$

where $\mathcal{X} \coloneqq \mathbb{R}^p$ denotes the state space of the system. Note that in the above hierarchy, by definition of r, we have $\mathcal{U}_r \neq \mathcal{U}_{r-1}$. It is equally straightforward to verify that $\mathcal{U}_1, \ldots, \mathcal{U}_r$ are subspaces.

The following lemma shows that in (4.4), the set containments are in fact strict, so that the $k-$step unsecurable subspaces are strictly decreasing in k.

Lemma 3. *Suppose that* $r > 1$. *Then,* $\forall i \leq r-1, \mathcal{U}_{i+1} \subset \mathcal{U}_i$.

Proof. The proof for the case when $i = r-1$ follows trivially by the very definition of r. Consider the case when $i \leq r-2$, and suppose for contradiction that for some $i \leq r-2, \mathcal{U}_i = \mathcal{U}_{i+1}$. Then, it follows from (4.3) that $\forall \mathbf{v} \in \mathcal{U}_{i+1},\ \exists \mathbf{d}$ such that $A\mathbf{v} + B^M\mathbf{d} \in \mathcal{U}_i = \mathcal{U}_{i+1}$. This implies, from the definition of V, that $\mathcal{U}_{i+1} \subseteq V$, which contradicts (4.4) since $V \subset \mathcal{U}_{i+1}$. □

The case $\mathcal{N}(C^H) = \mathcal{X}$ is of limited interest, as is made clear in the next section. We assume henceforth that this is not the case. Combining this assumption with Lemma 1, we have

$$V = \mathcal{U}_r \subset \mathcal{U}_{r-1} \subset \ldots \subset \mathcal{U}_1 \subset \mathcal{X}. \tag{4.5}$$

The above hierarchy is sufficient to assert the existence of orthonormal bases $\widetilde{B}_0, \widetilde{B}_1, \ldots, \widetilde{B}_r$ for $\mathcal{X}, \mathcal{U}_1, \ldots, \mathcal{U}_r$, respectively, with $\widetilde{B}_r \subset \widetilde{B}_{r-1} \subset \ldots \subset \widetilde{B}_0$. Now, using $\mathcal{X}, \mathcal{U}_1, \ldots, \mathcal{U}_r$, we derive $r+1$ orthogonal subspaces as follows.

Definition 6. *For* $k = 1, \ldots, r-1$, *define the* $k-$step securable subspace *as*

$$S_k \coloneqq \mathcal{U}_k \ominus \mathcal{U}_{k+1}, \tag{4.6}$$

where $\mathcal{U}_k \ominus \mathcal{U}_{k+1} \coloneqq$ *Span*$\{\widetilde{B}_k \backslash \widetilde{B}_{k+1}\}$. *The* $0-$step securable subspace *is defined as* $S_0 \coloneqq \mathcal{X} \ominus \mathcal{U}_1 = \mathcal{N}(C^H)^\perp$.

Intuitively, the $k-$step securable subspace is a set of states which when the system is initialized at, can be made indistinguishable to the honest sensors from the zero initial state for *exactly* k epochs by some choice of inputs of the malicious actuators. In particular, it cannot be made indistinguishable for $k+1$ epochs. A precise interpretation is given in Lemma 3.

The following lemma establishes the geometry of the securable subspace: The securable subspace is the span of all $k-$step securable subspaces.

Lemma 4. *The securable subspace* $S \coloneqq V^\perp$ *is composed of* $k-$*step securable subspaces in that* $S = S_0 + \ldots + S_{r-1}$.

Proof. $S_0 + \ldots + S_{r-1} = (\mathcal{X} \ominus \mathcal{U}_1) + (\mathcal{U}_1 \ominus \mathcal{U}_2) + \ldots + (\mathcal{U}_{r-1} \ominus V) = \mathcal{X} \ominus V = V^\perp$. □

Lemma 5. *For every* $\mathbf{s}_k \in S_k,\ \exists \mathbf{d}_0, \ldots, \mathbf{d}_{k-2}$ *such that*

$$C^H \mathbf{s}_k = 0,$$

$$C^H(A^{i-1}\mathbf{s}_k + A^{i-2}B^M\mathbf{d}_0 + \ldots + B^M\mathbf{d}_{i-2}) = 0,\ i = 2, \ldots, k.$$

Moreover, for every $0 \neq \mathbf{s}_k \in S_k,\ \nexists \mathbf{d}_0, \ldots, \mathbf{d}_{k-1}$ *such that* $C^H\mathbf{s}_k = 0,\ C^H(A^{i-1}\mathbf{s}_k + A^{i-2}B^M\mathbf{d}_0 + \ldots + B^M\mathbf{d}_{i-2}) = 0,\ i = 2, \ldots, k+1$.

Proof. The first statement of the lemma follows from the fact that $S_k \subset \mathcal{U}_k$. To prove the second statement of the lemma, suppose the contrary. Then, it follows that $\mathbf{s}_k \in \mathcal{U}_{k+1}$. Since $S_k \perp \mathcal{U}_{k+1}$ and $\mathbf{s}_k \neq 0$, we have $\mathbf{s}_k \notin S_k$, a contradiction. □

Lemma 6. *For every* $0 \neq \mathbf{s}_k \in S_k,\ k \in \{1, \ldots, r-1\}, \exists \mathbf{d}$ *such that* $A\mathbf{s}_k + B^M\mathbf{d} \in S_{k-1}$.

Proof. Omitted. □

4.4 Operational Meaning of the Securable Subspace for Stochastic Systems

In this section, we establish operational meanings for securable and unsecurable subspaces in the context of fully-observed stochastic linear dynamical systems of the form (4.1).

Observe that if there are no malicious nodes in the system, then, $\mathbf{x}[t+1] - A\mathbf{x}[t] - Bg_t(\mathbf{z}^t) = \mathbf{w}[t+1]$, and hence,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} [\mathbf{x}[t+1] - A\mathbf{x}[t] - Bg_t(\mathbf{z}^t)][\mathbf{x}[t+1] - A\mathbf{x}[t] - Bg_t(\mathbf{z}^t)]^T = \sigma_w^2 I.$$

Based on this observation, each honest node performs the following tests to detect the presence of malicious nodes:

- **Test:** Check if

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} (\mathbf{z}[t+1] - A\mathbf{z}[t] - Bg_t(\mathbf{z}^t))(\mathbf{z}[t+1] - A\mathbf{z}[t] - Bg_t(\mathbf{z}^t))^T = \sigma_w^2 I. \quad (4.7)$$

If the reported measurements $\{\mathbf{z}\}$ do not pass this test, then the honest nodes declare the presence of malicious nodes in the system.

Recall from Section 4.1 that the notation $(\mathbf{x})_Q$ denotes the projection of vector $\mathbf{x}$ on the subspace Q. Let P_k denote the projection matrix corresponding to the subspace $\sum_{j=1}^{k} S_j$. It follows from (4.7) that for $k = 0, \ldots, r-1$,

$$\begin{aligned}
&\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} (\mathbf{z}[t+1] - A\mathbf{z}[t] - Bg_t(\mathbf{z}^t))_{\sum_{j=1}^{k} S_j}^T (\mathbf{z}[t+1] - A\mathbf{z}[t] - Bg_t(\mathbf{z}^t))_{\sum_{j=1}^{k} S_j} \\
&= Tr[(\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} (\mathbf{z}[t+1] - A\mathbf{z}[t] - Bg_t(\mathbf{z}^t))^T P_k^T P_k (\mathbf{z}[t+1] - A\mathbf{z}[t] - Bg_t(\mathbf{z}^t))] \\
&= \sigma_w^2 Tr[P_k P_k^T]. \quad (4.8)
\end{aligned}$$

Defining $\sigma_k^2 \coloneqq \sigma_w^2 Tr[P_k P_k^T]$, from (4.8) we have that for $k = 0, \ldots, r-1$,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} (\mathbf{z}[t+1] - A\mathbf{z}[t] - Bg_t(\mathbf{z}^t))^T_{\sum_{j=1}^k S_j} (\mathbf{z}[t+1] - A\mathbf{z}[t] - Bg_t(\mathbf{z}^t))_{\sum_{j=1}^k S_j} = \sigma_k^2. \quad (4.9)$$

The following theorem establishes the operational meaning for the securable subspace in the context of fully observed stochastic linear dynamical systems.

Theorem 3. *Let* $\mathbf{m}[t] \coloneqq \mathbf{z}[t] - \mathbf{x}[t]$ *be the state estimation error of the honest nodes at time* t*. If the reported measurements* $\{\mathbf{z}\}$ *pass the test (4.7), then,*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} ||\mathbf{m}_S[t]||^2 = 0.$$

Proof. We show by induction on k that

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} ||\mathbf{m}_{S_k}[t]||^2 = 0$$

$\forall k \in \{0, \ldots, r-1\}$. I.e., the projection of the state estimation error on the $k-$step securable subspaces has zero power. This in turn implies the desired result.

We first note that since we have a system where all the states are directly observed, i.e., $C = I$, we have $S_0 = \mathcal{N}(C^H)^\perp = \mathit{Span}(\{\mathbf{e}_i : i \in \{o_1, \ldots, o_{h_S}\}\})$, where $\mathbf{e}_i$ is a $p \times 1$ vector whose i^{th} entry is unity and all other entries are zero. It follows that $\mathbf{m}_{S_0}[t] = \sum_{i=1}^{h_S} m_{o_i}[t]\mathbf{e}_{o_i} = 0$, where the last equality follows from the fact that at any time t, the distortion introduced by the honest sensors is zero. It follows that $\mathbf{m}_{S_0}[t] \equiv 0$, and so,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{m}_{S_0}^T[t]\mathbf{m}_{S_0}[t] = 0.$$

This serves as the base case for induction.

Assume for induction that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{m}_{S_i}^T[t]\mathbf{m}_{S_i}[t]=0$$

for $i=0,\ldots,q$. Since $\mathbf{z}[t+1]-A\mathbf{z}[t]-Bg_t(\mathbf{z}^t)=\mathbf{m}[t+1]-A\mathbf{m}[t]+B^M\mathbf{d}[t]+\mathbf{w}[t+1]$, substituting this in (4.9) for $k=q$ gives (recall from Section 4.1 the notation of $(\mathbf{x})_W^{TT}$)

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}(\mathbf{m}[t+1]-A\mathbf{m}[t]+B^M\mathbf{d}[t]+\mathbf{w}[t+1])_{\sum_{j=0}^{q}S_j}^{TT}=\sigma_q^2. \tag{4.10}$$

Now, write $\mathbf{m}[t]=\mathbf{m}_V[t]+\mathbf{m}_S[t]$, and $\mathbf{m}_S[t]=\mathbf{m}_{S_0}[t]+\ldots+\mathbf{m}_{S_{r-1}}[t]$. It follows from the definition of the unsecurable subspace that for all t, there exists $\mathbf{d}^V[t]$ such that $A\mathbf{m}_V[t]+B^M\mathbf{d}^V[t]\in V$, and from Lemma 4 that there exist $\mathbf{d}^{S_{q+2}}[t],...,\mathbf{d}^{S_{r-1}}[t]$ such that $(-A\mathbf{m}_{S_i}[t]+B^M\mathbf{d}^{S_i}[t])\in S_{i-1},\ i=q+2,\ldots,r-1$. Define $\mathbf{d}^U[t]\coloneqq\mathbf{d}[t]-\mathbf{d}^{S_{q+2}}[t]-\ldots-\mathbf{d}^{S_{r-1}}[t]-\mathbf{d}^V[t]$. Substituting these in (4.10) gives

$$\begin{aligned}
\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}(&\mathbf{m}_{S_0}[t+1]-A\mathbf{m}_{S_0}[t]\\
&+\mathbf{m}_{S_1}[t+1]-A\mathbf{m}_{S_1}[t]\\
&+\ldots\\
&+\mathbf{m}_{S_q}[t+1]-A\mathbf{m}_{S_q}[t]\\
&+\mathbf{m}_{S_{q+1}}[t+1]-A\mathbf{m}_{S_{q+1}}[t]+B^M\mathbf{d}^U[t]\\
&+\mathbf{m}_{S_{q+2}}[t+1]-A\mathbf{m}_{S_{q+2}}[t]+B^M\mathbf{d}^{S_{q+2}}[t]\\
&+\mathbf{m}_{S_{q+3}}[t+1]-A\mathbf{m}_{S_{q+3}}[t]+B^M\mathbf{d}^{S_{q+3}}[t]\\
&+\ldots\\
&+\mathbf{m}_{S_{r-1}}[t+1]-A\mathbf{m}_{S_{r-1}}[t]+B^M\mathbf{d}^{S_{r-1}}[t]\\
&+\mathbf{m}_V[t+1]-A\mathbf{m}_V[t]+B^M\mathbf{d}^V[t]\\
&+\mathbf{w}[t+1])_{\sum_{j=0}^{q}S_j}^{TT}=\sigma_q^2.
\end{aligned} \tag{4.11}$$

Since $-A\mathbf{m}_{S_i}[t] + B^M\mathbf{d}^{S_i}[t] \in S_{i-1}$ for $i = q+2, \ldots, r-1$, they vanish when projected on the subspace $\sum_{j=0}^{q} S_j$, i.e., their projection is the zero vector. Similarly, $\mathbf{m}_{S_{q+1}}[t+1], \ldots, \mathbf{m}_{S_{r-1}}[t+1]$ also vanish when projected on the subspace $\sum_{j=0}^{q} S_j$. Therefore, the above reduces to

$$\begin{aligned}\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} (&\mathbf{m}_{S_0}[t+1] - A\mathbf{m}_{S_0}[t] \\ &+ \mathbf{m}_{S_1}[t+1] - A\mathbf{m}_{S_1}[t] \\ &+ \ldots \\ &+ \mathbf{m}_{S_q}[t+1] - A\mathbf{m}_{S_q}[t] \\ &- A\mathbf{m}_{S_{q+1}}[t] + B^M\mathbf{d}^U[t] + \mathbf{w}[t+1])^{TT}_{\sum_{j=0}^{q} S_j} = \sigma_q^2. \end{aligned} \tag{4.12}$$

Note that from the induction hypothesis, we have that the term

$$\begin{aligned}\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} (&\mathbf{m}_{S_0}[t+1] - A\mathbf{m}_{S_0}[t] \\ &+ \mathbf{m}_{S_1}[t+1] - A\mathbf{m}_{S_1}[t] \\ &+ \ldots \\ &+ \mathbf{m}_{S_q}[t+1] - A\mathbf{m}_{S_q}[t])^{TT}_{\sum_{j=0}^{q} S_j} = 0. \end{aligned}$$

Expanding (4.12) and substituting the above equality yields

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} (-A\mathbf{m}_{S_{q+1}}[t] + B^M\mathbf{d}^U[t] + \mathbf{w}[t+1])^{TT}_{\sum_{j=0}^{q} S_j} = \sigma_q^2. \tag{4.13}$$

Using the martingale stability theorem [49, Lemma 2(iii)], it follows after some algebra that

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} (-A\mathbf{m}_{S_{q+1}}[t] + B^M\mathbf{d}^U[t])^{TT}_{\sum_{j=0}^{q} S_j} = 0. \tag{4.14}$$

Lemma 7. $\forall q \in \{1, \ldots, r-2\},\ \exists \epsilon > 0$ *such that* $\forall \mathbf{s}_{q+1} \in S_{q+1}$ *and for any* $\mathbf{d}$,

$$(A\mathbf{s}_{q+1} + B^M\mathbf{d})^{TT}_{\sum_{j=0}^{q} S_j} \geq \epsilon\, \mathbf{s}_{q+1}^T \mathbf{s}_{q+1}. \tag{4.15}$$

Proof. Let $\widehat{\mathbf{s}}_{q+1} \in S_{q+1}$ with $||\widehat{\mathbf{s}}_{q+1}|| = 1$. Define $I(\widehat{\mathbf{s}}_{q+1}) \coloneqq \inf_d ||(A\widehat{\mathbf{s}}_{q+1} + B^M\mathbf{d})_{\sum_{j=0}^{q} S_j}|| = \inf_d ||\widetilde{A}\widehat{\mathbf{s}}_{q+1} + \widetilde{B}^M\mathbf{d}||$, where $\widetilde{A} \coloneqq P_q A, \widetilde{B}^M \coloneqq P_q B^M$, and P_q is the projection matrix for the subspace $\sum_{j=0}^{q} S_j$. Clearly, the value $I(\widehat{\mathbf{s}}_{q+1})$ can be achieved by a vector $\mathbf{d} =: \widehat{\mathbf{d}}(\widehat{\mathbf{s}}_{q+1})$, not necessarily unique, all of whose entries are finite. It is also easy to verify that for any $k \in \mathbb{R}$,

$$I(k\widehat{\mathbf{s}}_{q+1}) = kI(\widehat{\mathbf{s}}_{q+1}), \tag{4.16}$$

and that $\widehat{\mathbf{d}}(k\widehat{\mathbf{s}}_{q+1}) = k\widehat{\mathbf{d}}(\widehat{\mathbf{s}}_{q+1})$.

Now, the function $I(\mathbf{x})$ is continuous in $\mathbf{x}$. To see this, let $\alpha(\mathbf{x}) \coloneqq \widetilde{A}\mathbf{x}$, and $\beta(\mathbf{x}) \coloneqq$ *dist*$(\mathbf{x}, \mathcal{R}(-\widetilde{B}^M))$ be the distance function to set $\mathcal{R}(-\widetilde{B}^M)$. Then, $I = \beta \circ \alpha$. Since both α and β are continuous functions, it follows that I is a continuous function.

Now, consider $\epsilon' = \inf_{||\mathbf{s}_{q+1}||=1} I(\mathbf{s}_{q+1})$, which is the minimization of a continuous function over a compact set $\Delta \coloneqq \{\mathbf{s} \in S_{q+1} :\ ||\mathbf{s}|| = 1\}$. It follows that the infimum is attained by some element $\mathbf{s}^*_{q+1} \in \Delta$, i.e., $\epsilon' = ||(A\mathbf{s}^*_{q+1} + B^M\mathbf{d}^*)_{\sum_{j=0}^{q} S_j}||$ for some $\mathbf{d}^*$. Note that since $0 \neq \mathbf{s}^*_{q+1} \in S_{q+1}$, we have $(A\mathbf{s}^*_{q+1} + B^M\mathbf{d})_{\sum_{j=0}^{q} S_j} \neq 0$ for all choices of $\mathbf{d}$, and hence, $\epsilon' > 0$. Combining this with (4.16), we have

$$\begin{aligned} ||(A\mathbf{s}_{q+1} + B^M\mathbf{d})_{\sum_{j=0}^{q} S_j}|| = ||(A||\mathbf{s}_{q+1}||\frac{\mathbf{s}_{q+1}}{||\mathbf{s}_{q+1}||} + B^M\mathbf{d})_{\sum_{j=0}^{q} S_j}|| &= I(||\mathbf{s}_{q+1}||\frac{\mathbf{s}_{q+1}}{||\mathbf{s}_{q+1}||}) \\ &= ||\mathbf{s}_{q+1}||I(\frac{\mathbf{s}_{q+1}}{||\mathbf{s}_{q+1}||}) \geq ||\mathbf{s}_{q+1}||\epsilon'. \end{aligned}$$

Squaring the above inequality completes the proof. □

Combining (4.15) with (4.14), we have

$$0 = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} ||(-A\mathbf{m}_{S_{q+1}}[t] + B^M \mathbf{d}^U[t])_{\sum_{j=0}^{q} S_j}||^2 \geq \epsilon \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} ||(\mathbf{m}_{S_{q+1}}[t])_{\sum_{j=0}^{q} S_j}||^2,$$

which completes the proof. □

The following corollary shows that the unsecurable subspace is, in general, the minimal subspace "containing" all of the covariance of the state estimation error.

Corollary 1. *Under the same conditions as in Theorem 1,*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{m}_S[t]\mathbf{m}_S^T[t] = 0. \tag{4.17}$$

Consequently,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{m}[t]\mathbf{m}^T[t] = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{m}_V[t]\mathbf{m}_V^T[t]. \tag{4.18}$$

Proof. Let $m_{S,i}[t]$ denote the i^{th} entry of $\mathbf{m}_S[t]$. Then, it follows from Theorem 1 that

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} m_{S,i}^2[t] = 0$$

for $i = 1, \ldots, p$. It follows that for all i, j,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} m_{S,i}[t]m_{S,j}[t] = 0.$$

□

The above results establish the extent to which malicious sensors can distort the actual measurements in a stealthy attack. Given a control objective, this also has implications on the extent to which malicious actuators can deviate from their control law-specified inputs, and consequently, the extent to which the actual performance metrics of the system deviate from their optimal values.

4.5 Concluding Remarks

The notions of securable and unsecurable subspaces of a linear dynamical system, introduced in the context of secure control, have important operational meanings for deterministic linear dynamical systems. This chapter has examined these subspaces from the vantage point of a fully-observed stochastic linear dynamical system, and has established operational meanings for them in the stochastic context. Specifically, the securable subspace is shown to be, in general, the maximal subspace of the state space on which the projection of the state estimation error incurred by the honest nodes, due to the presence of malicious nodes, has zero covariance. Consequently, the unsecurable subspace is, in general, the minimal subspace that "contains" all of the covariance of the state estimation error. Extensions of this result to more general models such as that of partially-observed stochastic linear dynamical systems with both process and observation noise are reported in the next chapter.

5. ON THE OPERATIONAL SIGNIFICANCE OF THE SECURABLE SUBSPACE FOR PARTIALLY OBSERVED LINEAR STOCHASTIC SYSTEMS*

5.1 Introduction

In this chapter, we generalize the results in Chapters 3 and 4 and establish the operational meaning of the securable subspace in the context of general multiple-input-multiple-output partially observed stochastic linear dynamical systems with both Gaussian process and observation noise, which class constitutes one of the most general models used in modern linear system theory. Specifically, we show that for these systems, the securable subspace has the operational meaning as that subspace of the state space along which the malicious sensors and actuators cannot distort the MMSE state estimate that would be obtained had there been no malicious nodes in the system, beyond adding a zero power sequence, if they wish to remain undetected. This in turn implies that the covariance of the projection of the state estimation error of the honest nodes on the securable subspace remains at its designed value in spite of *arbitrary* attack strategies of the malicious nodes, if the malicious activity is to remain undetected.

The rest of the chapter is organized as follows. Section 5.2 describes the problem setting that we consider in this chapter. Section 5.3 is devoted to establishing the operational meaning of the securable subspace in the context of general partially-observed stochastic linear dynamical systems. Section 5.4 concludes the chapter.

Notation: Given a vector x, we denote by x^{TT} the outer product xx^T. Given subspaces $\mathcal{W}_1, \ldots, \mathcal{W}_p$ of some vector space, we denote by $\oplus_{l=1}^{p} \mathcal{W}_l$ the span of subspaces $\mathcal{W}_1, \ldots, \mathcal{W}_p$. Given a vector x and a subspace $\mathcal{W}$, we denote by $x_{\mathcal{W}}$ the projection of x on $\mathcal{W}$.

*Reprinted with permission from "On the Operational Significance of the Securable Subspace for Partially Observed Linear Stochastic Systems," by B. Satchidanandan and P. R. Kumar in *Proceedings of the 2018 IEEE Conference on Decision and Control (CDC),* pp. 2068-2073, 2018, IEEE.

5.2 Problem Setup

Consider a p^{th} order linear dynamical system with m inputs and n outputs described by

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) + w(t+1), \\ y(t+1) &= Cx(t+1) + n(t+1), \end{aligned} \tag{5.1}$$

where $x(t) \in \mathbb{R}^p$, $u(t) \in \mathbb{R}^m, y(t) \in \mathbb{R}^n$, $w(t) \sim \mathcal{N}(0, Q)$ is the process noise distributed i.i.d. across time, $n(t) \sim \mathcal{N}(0, R)$ is the measurement noise distributed i.i.d. across time, the process and measurement noise processes are independent, and A, B, C, Q, R are matrices of appropriate dimensions. We further assume that (A, C) is observable, $(A, Q^{1/2})$ is reachable, and $R > 0$.

As throughout this thesis, the scenario that we address in this chapter is one where an arbitrary subset of the sensors and actuators in the system could be malicious. A malicious sensor need not report the measurements that it observes truthfully, and a malicious actuator need not apply inputs in accordance with the designated control policy. The honest nodes in the system do not know the identity of the malicious nodes in the system, or even if there are any malicious nodes in the system or not. The malicious nodes, however, know the identity of all other malicious nodes in the system, so that they can exchange information among themselves thereby allowing them to carry out coordinated attacks.

We assume that an arbitrary, possibly history-dependent control policy $\{g_t\}$ is in place which specifies the control input $u^g(t)$ that has to be applied at time t as a function of the reported measurements $z^t := (z(0), \ldots, z(t))$ upto time t, i.e., $u^g(t) = g_t(z^t)$. If actuator i, $i \in \{1, \ldots, m\}$, is honest, then $u_i(t) \equiv u_i^g(t)$ for all t. However, if actuator-i is malicious, then this need not be the case, and the actuator can apply any input that it wishes. We denote by $d(t)$ the "input distortion" introduced by the malicious actuators, i.e., $d(t) := u^M(t) - u^{gM}(t)$, where $u^M(t)$ is a vector that collects the entries of $u(t)$ corresponding to the malicious actuators, and $u^{gM}[t]$ is a vector that collects the control policy-specified inputs corresponding to the malicious actuators. Clearly, these vectors belong to $\mathbb{R}^{m_a}$, where m_a is the number of malicious actuators in the system.

Consequently, the system (5.1) evolves in closed loop according to

$$\begin{aligned} x(t+1) &= Ax(t) + Bu^g(t) + B^M d(t) + w(t+1), \\ y(t+1) &= Cx(t+1), \end{aligned} \tag{5.2}$$

where B^M denotes the submatrix of B formed by the columns corresponding to the malicious actuators.

In the previous chapters, we considered special cases of system (5.1), viz., the case of deterministic systems where there is no process or measurement noise, and the only uncertainty is in the initial conditions of the system [48], and the case of fully-observed stochastic systems (i.e., $C = I_p$ and there is no measurement noise) [50], and established operational meanings for the securable subspace in these contexts. In this chapter, we generalize the above results and establish the operational meaning of these subspaces in the general context of partially observed linear stochastic systems of the form (5.1) with both process and measurement noise, which is one of the most general models describing linear systems.

5.3 Operational Meaning of the Securable Subspace for Partially Observed Systems

This section is devoted to establishing the operational meaning of the securable subspace for partially observed linear stochastic systems.

Consider the linear dynamical system (5.1) evolving in closed loop according to (5.2). The MMSE-optimal state prediction $\widehat{x}^R(k+1|k)$, given the *true* actuation inputs and sensor measurements, is obtained from the Kalman filtering equations as

$$\begin{aligned} \widehat{x}^R(k+1|k) &= A\widehat{x}^R(k|k) + Bu^g(k) + B^M d(k), \\ \widehat{x}^R(k+1|k+1) &= \widehat{x}^R(k+1|k) + L\nu^R(k+1). \end{aligned} \tag{5.3}$$

where $\nu^R(k+1) \coloneqq y(k+1) - C\widehat{x}^R(k+1|k)$ is the "real" innovations at time k, and L is the steady-

state Kalman gain. For simplicity, we are assuming here that the system is in steady state and we are employing the steady-state Kalman gain. The proof extends in a straightforward manner to the use of time-varying gain, and is omitted for brevity of exposition. Note that the honest nodes in the system may not be able to compute $\widehat{x}^R(k+1|k)$ or $\widehat{x}^R(k+1|k+1)$ since they may not be privy to the true measurements $\{y\}$ or the input distortion $\{d\}$.

The honest nodes in the system, in the absence of any evidence on the contrary, begin with the presumption that all nodes in the system are honest, and based on this, compute the state estimates $\widehat{x}^F(k+1|k+1)$ as

$$\begin{aligned} \widehat{x}^F(k+1|k) &= A\widehat{x}^F(k|k) + Bu^g(k), \\ \widehat{x}^F(k+1|k+1) &= \widehat{x}^F(k+1|k) + L\nu^F(k+1). \end{aligned} \tag{5.4}$$

where $\nu^F(k+1) \coloneqq z(k+1) - C\widehat{x}^F(k+1|k)$. In order to check whether there are any malicious nodes in the system or not, the honest nodes conduct the following test.

Test: For all $q \in \{1, \ldots, r-1\}$, check if the reported sequence of measurements $\{z\}$ satisfy (recall the notation $(\cdot)^{TT}$ from Section 5.1)

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (z(k+q) - C\widehat{x}^F(k+q|k))^{TT} = \Sigma^q, \tag{5.5}$$

where $\Sigma^q \coloneqq \mathbb{E}[(y(k+q) - C\widehat{x}^R(k+q|k)^{TT}]$ is the covariance matrix of the MMSE-optimal $q-$step-ahead output prediction error. Note that if there are no malicious nodes in the system, the above test would be passed by the reported measurements almost surely, and therefore, if the reported measurements fail the above test, then the honest nodes can conclude the presence of malicious nodes in the system. Therefore, if the malicious nodes wish to remain undetected, then they are constrained to reporting measurements that pass (5.5). While the above test is presented in an asymptotic form, it can be converted into a finite-time test with certain false alarm and detection rates using standard methods.

The following theorem establishes the operational significance of the securable subspace in the context of partially observed stochastic linear dynamical systems.

Theorem 4. *Consider the system (5.1) with malicious sensors and actuators, and suppose that the pair (A, C) is observable, $(A, Q^{1/2})$ is reachable, and $R > 0$. If the reported measurements $\{z\}$ satisfy (5.5), then,*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} \left\| (\widehat{x}^F(k+1|k) - \widehat{x}^R(k+1|k))_{\mathcal{S}} \right\|^2 = 0.$$

Proof. Define

$$\delta(k+q|k) \coloneqq \widehat{x}^F(k+q|k) - \widehat{x}^R(k+q|k). \tag{5.6}$$

Since the reported measurements pass test (5.5), we have that for every $q \in \{1, \ldots, r-1\}$

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (z(k+q) - C\widehat{x}^F(k+q|k))^{TT} = \Sigma^q.$$

Let h_S be the number of honest sensors in the system, and assume without loss of generality that the sensors measuring the outputs y_1 through y_{h_s} are honest (since the rows of C and the corresponding entries of y, etc., can be rearranged to achieve this). We note that the upper h_s entries of $z(k+1)$, denoted $z^H(k+1)$, agree with the upper h_s entries of $y(k+1)$, denoted $y^H(k+1)$, since these h_s sensors are honest. Hence, equating the upper-left $h_s \times h_s$ sub-matrix of the above equality yields

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (y^H(k+q) - C^H\widehat{x}^F(k+q|k))^{TT} = \Sigma_H^q,$$

where Σ_H^q is the upper-left $h_s \times h_s$ sub-matrix of Σ^q. Using (5.6), the above can be written as

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (y^H(k+q) - C^H\widehat{x}^R(k+q|k) - C^H\delta(k+q|k))^{TT} = \Sigma_H^q,$$

Denoting the vector $C^H\delta(k+q|k)$ by $\delta^H(k+1|k)$, and the vector $y^H(k+q) - C^H\widehat{x}^R(k+q|k)$ by $\widetilde{y}^H(k+q|k)$, and noting that $\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}(\widetilde{y}^H(k+q|k))^{TT} = \Sigma_H^q$, the above equality simplifies to

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}\delta^H(k+q|k)\delta^{H^T}(k+q|k) - \widetilde{y}^H(k+q|k)(\delta^H(k+q|k))^T \\ -\delta^H(k+q|k)(\widetilde{y}^H(k+q|k))^T = 0. \tag{5.7}$$

Now, the second term in the above sum can be rewritten as

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}\widetilde{y}^H(k+q|k)\delta^H(k+q|k)^T = \sum_{i=0}^{q-1}\lim_{\tau\to\infty}\frac{1}{\tau}\sum_{g=1}^{\tau}\widetilde{y}^H(i+gq|i+(g-1)q) \\ \delta^H(i+gq|i+(g-1)q).$$

Now, $\widetilde{y}^H(i+gq|i+(g-1)q)$ is independent of $\mathcal{F}_{i+(g-1)q}$ whereas $\delta^H(i+gq|i+(g-1)q) \in \mathcal{F}_{i+(g-1)q}$. Moreover, $\{\widetilde{y}^H(i+gq|i+(g-1)q)\delta^H(i+gq|i+(g-1)q)\}$, viewed as a sequence in g, is i.i.d. across time. Using the Martingale Stability Theorem (MST) [49], (5.7) reduces to

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}\delta^H(k+q|k)\delta^{H^T}(k+q|k) = 0. \tag{5.8}$$

Now, note that

$$\widehat{x}^R(k+q|k) = A^{q-1}\widehat{x}^R(k+1|k) + \sum_{l=0}^{q-2}A^lBu^g(k+q-l-1|k) \\ -\sum_{l=0}^{q-2}A^lB^M\widehat{d}(k+q-l-1|k), \tag{5.9}$$

and that

$$\widehat{x}^F(k+q|k) = A^{q-1}\widehat{x}^F(k+1|k) + \sum_{l=0}^{q-2}A^lBu^g(k+q-l-1|k), \tag{5.10}$$

where $\widehat{d}(k+q-l-1|k) \coloneqq \mathbb{E}[d(k+q-l-1)|\mathcal{F}_k)]$. Subtracting (5.9) from (5.10) and left-multiplying the result by C^H yields

$$C^H\delta(k+q|k) = C^H[A^{q-1}\delta(k+1|k) - \sum_{l=0}^{q-2} A^l B^M \widehat{d}(k+q-l-1|k)].$$

Since the sequence $\{C^H\delta(k+q|k)\}$ is of zero power following (5.8) for all $q \in \{1,\dots,r\}$, we have

$$\lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} \left|\left|C^H[A^{q-1}\delta(k+1|k) - \sum_{l=0}^{q-2} A^l B^M \widehat{d}(k+q-l-1|k)]\right|\right|^2 = 0. \tag{5.11}$$

Now, in order to show that $\{\delta_{\mathcal{S}}(k+1|k)\}$ is of zero power, we show, via induction, that each of the sequences $\{\delta_{\mathcal{S}_0}(k+1|k)\},\dots,\{\delta_{\mathcal{S}_{r-1}}(k+1|k)\}$ is of zero power.

Note that substituting the value $q=1$ in (5.8) gives $\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}\left\|C^H\delta(k+1|k)\right\|^2 = 0$, which in turn implies that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}\left\|\delta_{\mathcal{S}_0}(k+1|k)\right\|^2 = \lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}\left\|C^{H^T}(C^HC^{H^T})^{-1}C^H\delta(k+1|k)\right\|^2 = 0.$$

This establishes the base case for induction.

Now, we assume for induction that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}\left\|\delta_{\mathcal{S}_i}(k+1|k)\right\|^2 = 0$$

for all $i \in \{0,\dots,j\}$. Substituting $q=j+1$ in (5.11) and expanding $\delta(k+1|k)$ in terms of its components in $\mathcal{S}_0,\dots,\mathcal{S}_{r-1}$ and V gives

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}\left|\left|C^H[\sum_{f=0}^{r-1}A^j\delta_{\mathcal{S}_f}(k+1|k) + A^j\delta_V(k+1|k) - \sum_{l=0}^{j-1}A^lB^M\widehat{d}(k+j-l|k)]\right|\right|^2 = 0.$$

Using the induction hypothesis, the above reduces to

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}\Big|\Big|C^H[\sum_{f=j+1}^{r-1}A^j\delta_{\mathcal{S}_f}(k+1|k)+A^j\delta_V(k+1|k)-\sum_{l=0}^{j-1}A^lB^M\widehat{d}(k+j-l|k)]\Big|\Big|^2=0. \tag{5.12}$$

Now, it follows from the definition of the unsecurable subspace that there exist $\widehat{d}^V(k+j|k),\ldots,\widehat{d}^V(k+1|k)$ such that

$$A^j\delta_V(k+1|k)-\sum_{l=0}^{j-1}A^lB^M\widehat{d}^V(k+j-l|k)\in V\subseteq\mathcal{N}(C^H). \tag{5.13}$$

Similarly, it follows from Lemma 1 that for every $f\in\{j+2,\ldots,r-1\}$, there exist $\widehat{d}^f(k+1|k),\ldots,\widehat{d}^f(k+j|k)$ such that

$$A^j\delta_{\mathcal{S}_f}(k+1|k)-\sum_{l=0}^{j-1}A^lB^M\widehat{d}^f(k+j-l|k)\in\mathcal{N}(C^H). \tag{5.14}$$

Consequently, define for $l=0,\ldots,j-1$,

$$\gamma(k+j-l|k)\coloneqq\widehat{d}(k+j-l|k)-\widehat{d}^V(k+j-l|k)-\sum_{f=j+2}^{r-1}\widehat{d}^f(k+j-l|k).$$

Substituting for $\widehat{d}(k+j-l|k)$ in (5.12) using the above equality and then applying (5.13) and (5.14) on the result gives

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}\Big|\Big|C^H[A^j\delta_{\mathcal{S}_{j+1}}(k+1|k)-\sum_{l=0}^{j-1}A^lB^M\gamma(k+j-l|k)]\Big|\Big|^2=0. \tag{5.15}$$

Using Lemma 1 and Lemma 2, the above yields, after some algebra, that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}\big\|\delta_{\mathcal{S}_{j+1}}(k+1|k)\big\|^2=0,$$

thereby completing the induction. □

Let $\widetilde{x}^F(k+1|k) \coloneqq x(k+1) - \widehat{x}^F(k+1|k)$, and define $\widetilde{x}^R(k+1|k)$ likewise. The following result presents an equivalent operational meaning for the securable subspace.

Corollary 2. *Assume as in Theorem 1. If the reported measurements $\{z\}$ satisfy the test (5.5), then,*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} \widetilde{x}^F_{\mathcal{S}}(k+1|k)(\widetilde{x}^F_{\mathcal{S}}(k+1|k))^T = \lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} \widetilde{x}^R_{\mathcal{S}}(k+1|k)(\widetilde{x}^R_{\mathcal{S}}(k+1|k))^T.$$

It follows from this that the covariance of the projection of the state estimation error of the honest nodes on the securable subspace remains at its optimal value.

Proof. Note that

$$\begin{aligned}
\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} \widetilde{x}^F_{\mathcal{S}}(k+1|k)(\widetilde{x}^F_{\mathcal{S}}(k+1|k))^T &= \lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (x_{\mathcal{S}}(k+1) - \widehat{x}^F_{\mathcal{S}}(k+1|k))^{TT} \\
&= \lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (x_{\mathcal{S}}(k+1) - \widehat{x}^R_{\mathcal{S}}(k+1|k) - \delta_{\mathcal{S}}(k+1|k))^{TT} \\
&= \lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (x_{\mathcal{S}}(k) - \widehat{x}^R_{\mathcal{S}}(k+1|k))^{TT}, \qquad (5.16)
\end{aligned}$$

where the last equality follows as a consequence of Theorem 1. □

5.4 Concluding Remarks

In this chapter, we have considered multiple-input-multiple-output partially observed linear stochastic systems with both process and measurement noise, in which an arbitrary subset of the sensors and actuators could be malicious. We have shown that for any such system, its state space can be decomposed into two orthogonal subspaces, called the securable and the unsecurable subspaces, such that the covariance of the projection of the one-step-ahead state estimation error of the honest nodes on the securable subspace remains at its optimal value regardless of what attack strategy the malicious sensors and actuators choose to employ, as long as the malicious sensors

and actuators wish to remain undetected. The malicious nodes, therefore, can degrade the state estimation performance only along the unsecurable subspace of the linear dynamical system.

6. DYNAMIC WATERMARKING: ACTIVE DEFENSE OF NETWORKED CYBERPHYSICAL SYSTEMS*

6.1 Introduction

Fig. 6.1, reproduced from Chapter 1, illustrates the basic architecture of a Networked Cyber-Physical System. At the heart of the system is a physical plant with m inputs and n outputs. Each input is controlled by an independent actuator, and each output is measured by an independent sensor. The dynamics of the physical plant along with the ambient noises map the actuation signals applied by the actuators to outputs that are measured by the sensors. The plant model and the noise statistics are known to all sensors and actuators in the system. The outputs observed by the sensors are communicated to entities called controllers through an underlying communication network. Each controller's job is to compute, in accordance with a specified control policy, the particular actuation signals that must be applied. The result of this computation is then communicated to the actuators through the communication network, which then apply the actuation signals. We assume that the network is complete, so that every node in the network can communicate with every other node. We use the term "nodes" generically to refer to any entity in the network. Therefore, in Fig. 6.1, some nodes are actuators, some are sensors, some are controllers, and some could just be relays whose only job is to forward the information from one node to another. In this work, we assume that each node has both communication and computational capabilities, thereby allowing the controllers to be collocated with the actuators.

However, certain sensors in the system could be "malicious" (suggestive of this, some nodes in Fig. 6.1 are marked in red), and the other nodes are said to be "honest." We further assume that the malicious nodes know the identity of all other malicious nodes in the system, allowing them to

*All sections of this chapter except Section 6.9 are reprinted with permission from "Dynamic Watermarking: Active Defense of Networked Cyber-Physical Systems," by B. Satchidanandan and P. R. Kumar in *Proceedings of the IEEE,* vol. 105, no. 2, pp. 219-240, Feb. 2017, IEEE.

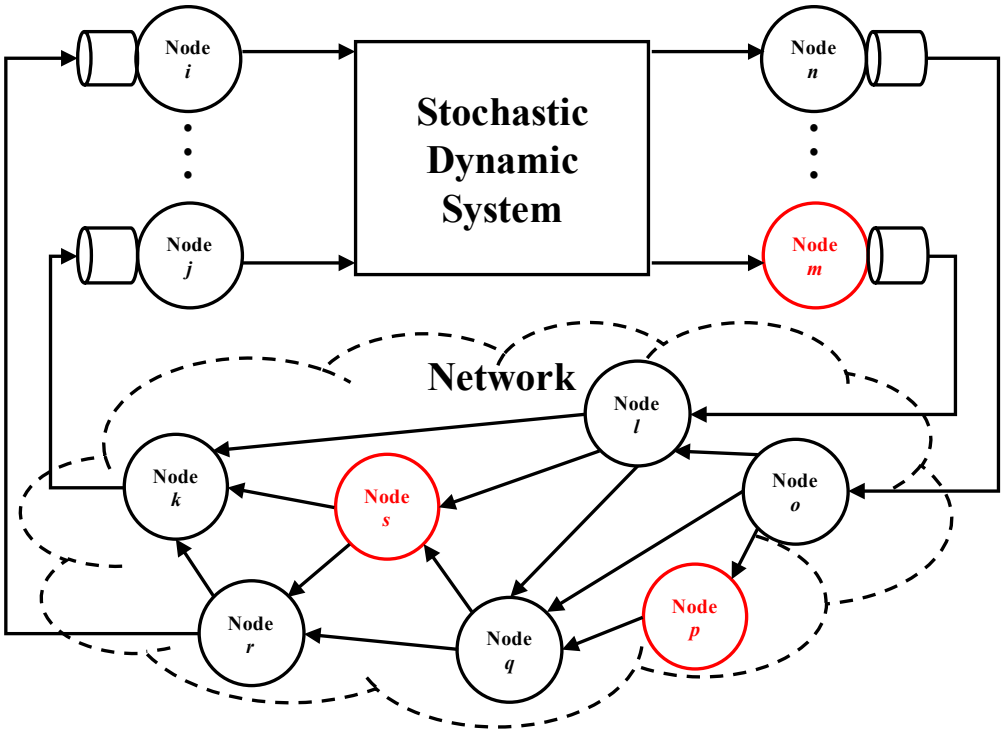

Figure 6.1: A Networked Cyber-Physical System

collude to achieve their objective, whereas the honest nodes don't know which of the other nodes are malicious or honest. A malicious sensor is a sensor node which does not report accurately the measurements that it observes. Rather, it reports a distorted version of the measurements. A malicious router node may not forward the packets that it receives, may forward packets that it does not receive (while claiming otherwise), alter packets before forwarding, introduce intentional delays, impersonate some other node in the system, etc. Therefore, the challenges of securing a Networked Cyber-Physical System are two-fold:

1) Secure the cyber layer that comprises the communication network, ensuring confidentiality, integrity, and availability of network packets, and

2) Secure the sensors and actuators interfacing with the physical layer.

The former is achieved by a combination of traditional approaches such as cryptography and a more recent line of work reported in [29, 51, 52], while the latter is the subject of this chapter.

Based on our assumption that the communication network is complete, and that the cyber layer has been secured, we suppose that every node in the network knows the identity of the node from which a packet that it receives originated, and knows if a packet that it receives was tampered with by any node along the route. Therefore, in the sequel, we abstract the cyber layer as consisting of secure, reliable, delay guaranteed bit pipes between any pair of nodes in the network. In particular, we imagine that there exists a secure, reliable, delay guaranteed bit pipe between any particular sensor and actuator node.

Finally, the plant that is being controlled is abstracted as a stochastic linear dynamical system described by time-invariant parameters. While any system that is of practical interest is most certainly non-linear, we focus on linear systems for two reasons. The first is that a linear system lends itself to tractable analysis, and enables one to separate the complexity arising out of the problem at hand from the complexity arising as a consequence of the system's non-linearity. Secondly, a theory developed for linear systems provides valuable insights and design principles that often transcend the particulars of the model and apply to a much broader class of systems. The wide applicability of Kalman's pioneering work on linear control systems [53] stands testimony to this fact.

This chapter is organized as follows. Section 6.2 provides a system-theoretic formulation of the problem. Section 6.3 describes our approach of active defense for networked cyber-physical systems. Section 6.4 opens by describing the method in the relatively simple context of a scalar linear Gaussian system and rigorously establishes the associated theoretical guarantees. Section 6.5 treats the more general class of scalar auto-regressive systems with exogenous noise (ARX systems) that is Gaussian. Section 6.6 extends these ideas to the more general ARMAX systems with arbitrary delay, a model that is frequently encountered in process control. Section 6.7 deals with partially observed SISO systems with Gaussian process and measurement noise. Section 6.8 considers perfectly observed multi-input, multi-output linear Gaussian systems in state-space form, and Section 6.9 treats the general case of a partially observed MIMO system with both process and measurement noise. Section 6.10 describes how our results can potentially be extended to non-

Gaussian systems. Section 6.11 shows how the theoretical results lead to statistical tests that can be used to detect malicious behavior within a delay bound with a controlled false alarm rate.

6.2 Problem Formulation

We are now in a position to state the problem in precise terms (with notation as indicated in the Appendix). Consider an $m \times n$ stochastic linear dynamical system of order p, described by

$$\mathbf{x}[t+1] = A\mathbf{x}[t] + B\mathbf{u}[t] + \mathbf{w}[t+1], \tag{6.1}$$

where $A \in \mathbb{R}^{p\times p}$, $B \in \mathbb{R}^{p\times m}$, $\mathbf{u}[t]$ is the input applied to the plant at time t, $\{\mathbf{w}\}$ is a sequence of independent and identically distributed (i.i.d.) Gaussian random vectors with zero mean and covariance martix $\sigma_w^2 I$, independent of the initial state $\mathbf{x}_0$ of the system.

While a malicious sensor can report different measurements to different actuators, the consistency of the reported measurements can be checked by allowing the actuators to exchange the reported measurements among themselves. This constrains the malicious sensors to report the same value to all honest nodes in the system. The sensor node j reports to the honest nodes in the system the value $z_j[t]$ as the measurement that it observes at time t. We define $\mathbf{z}[t] \coloneqq [z_1[t]\ z_2[t]\ z_3[t] \cdots z_n[t]]^T$. We will call $\mathbf{z}[t]$ for $t \geq 0$ the measurements *reported* by the sensors. Note that the sensor j is honest if $z_j[t] = x_j[t]\ \forall t$.

We assume that a control policy is in place, known to all nodes in the system, and allow for it to be history dependent, so that the i^{th} input at time t, $u_i^g[t]$, dictated by the policy is

$$u_i^g[t] = g_t^i(\mathbf{z}^t), \tag{6.2}$$

where $\mathbf{z}^t \coloneqq \{\mathbf{z}[0], \mathbf{z}[1], ..., \mathbf{z}[t]\}$. We can suppose without loss of generality that the controller that computes this control law is collocated with the actuator node.

Our goal is to secure the control system by developing techniques that prevent the malicious nodes from causing excessive distortion if they are to remain undetected. We will suppose that the

purpose of control law (6.2) is to improve the regulation performance of the system with respect to the disturbances affecting it. For some of the performance results we will assume that the system is open-loop stable, and show that the adversarial nodes cannot affect the performance of the system without remaining undetected, and if detected then they can be disconnected, returning the system to stable behavior. The fundamental results characterizing what is the most that the adversarial nodes can do while remaining undetected, are applicable to all systems, stable or unstable.

6.3 Dynamic Watermarking

The key idea which allows the honest nodes to detect the presence of malicious nodes in the system is the following. Let g denote the control policy in place that specifies inputs to be applied in response to observed outputs. At each time instant t, an actuator node superimposes on its control policy-specified input $u_i^g[t]$, a random variable $e_i[t]$ that it draws independently from a specified distribution. Therefore, the input that actuator i applies at time t is

$$u_i[t] = u_i^g[t] + e_i[t]. \tag{6.3}$$

This is illustrated in Fig. 6.2. The random variables $e_0[t], e_1[t], e_2[t] \cdots$ are independent and identically distributed (i.i.d.), independent of the control policy-specified input. The distribution that they are chosen from is made public (i.e., made known to every node in the system), but the actual values of the excitation are not disclosed. In fact, this is how the honest actuators can check whether signals are being tampered with as they travel around the control loop. We refer to these random variables as an actuator node's *privately imposed excitation*, since only that actuator node knows the actual realization of the sequence.

To see why private excitation helps, consider the example of a single-input-single-output (SISO) system where the sensor is malicious and the actuator is honest. Suppose that the control policy in place is $g = (g_1, g_2, \ldots, g_t, \ldots)$, where g_t specifies the input to be applied at time t in response to outputs up till that time. The actual outputs of the plant up to time t are $x^t := (x[0], x[1], \ldots, x[t])$. However the outputs reported to it by the malicious sensor are $z^t := (z[0], z[1], \ldots, z[1])$, which

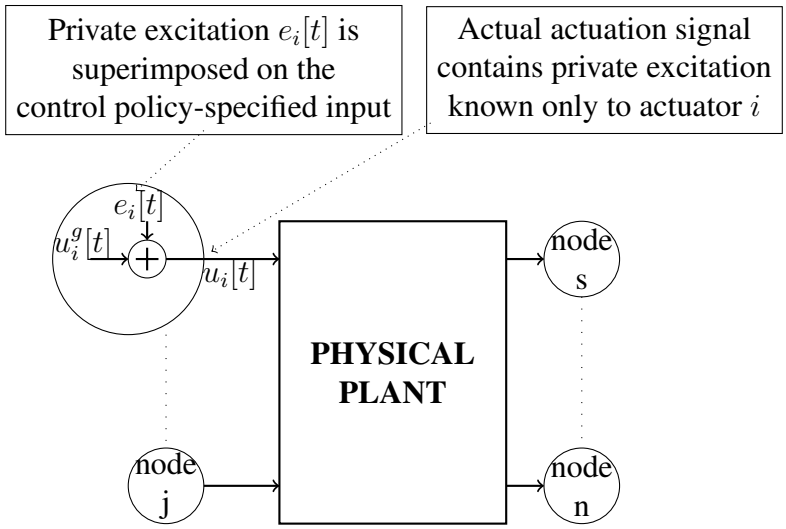

Figure 6.2: The actuator node i superimposes a private excitation whose realization is unknown to other nodes on to its control inputs. Reprinted with permission from [1].

may differ from x^t. The actuator therefore applies the control input $u[t] = g_t(z^t)$ at time t, without using any private excitation. Then, we have the system

$$x[t+1] = ax[t] + bg_t(z^t) + w[t+1], \tag{6.4}$$

where $w[t]$ is the zero-mean process noise with variance σ_w^2.

If the actuator does not superimpose a private excitation, the sensor knows, for each time t, the input $u[t]$ applied by the actuator. This is because it knows both the measurement sequence z^t that it reported to the actuator as well as the control policy $\{g_1, g_2, ...\}$ that the honest actuator has implemented. Hence, the malicious sensor can report a sequence of measurements $\{z[t]\}$ to the actuator *without even "looking" at the output*, but by simply "simulating a linear stochastic system" after generating *its own i.i.d. process noise* $\{w'\}$ from the same distribution as that of $\{w\}$, as follows.

$$z(t+1) = az(t) + bg_t(z^t) + w'(t+1). \tag{6.5}$$

The actuator cannot detect that the sensor is malicious since the sequence $\{w'\}$, having been chosen from the same distribution as $\{w\}$, could have been the actual process noise.

However, by superimposing a private excitation that is *unknown* to the sensor, the actuator *forces* the sensor to report measurements that are correlated with $\{e_i\}$, lest it be exposed, as we will show in the sequel. In the following sections, we prove that thereby constraining the sensor to report measurements that are correlated with the private excitation essentially limits the amount of distortion that the sensor can get away with while remaining undetected to be essentially zero in a mean-square sense.

This active defense technique is similar in spirit to the technique of digital watermarking [10] in electronic documents. Electronic documents can be easily transmitted and reproduced in large numbers. In doing so, the source of the document may be deleted, and can result in copyright violations. To protect the identity of its author, or any other information about the document, the electronic document is "watermarked" before being made available electronically. A digital watermark is a digital code that is robustly embedded in the original document [11]. By robust, it is meant that the code cannot be destroyed without destroying the contents of the document. This code typically contains information about the document that needs to be preserved. Though preferable, it is not a requirement that the watermark be imperceptible. The only requirement is that it does not distort the actual contents beyond certain acceptable limits [11]. Applications of digital watermarking also include data authentication, where fragile watermarks are used which get destroyed when the data is tampered with [11], and data monitoring and tracking.

This approach for secure control is analogous to digital watermarking. As we show in the subsequent sections, with regard to, for example, the above case where we considered a compromised sensor, injecting private excitation that is unknown to the sensor effectively "watermarks" the process noise, in the sense that the sensor cannot separate the private excitation from the process noise. Hence, any attempt on the part of the sensor to distort the process noise (which is the only component of the output unknown to the actuators) will also distort the "watermark," allowing the honest nodes to detect malicious activity.

Specifically, given a general MIMO system of the form (6.1), we define

$$\mathbf{v}[k] \coloneqq \mathbf{z}[k] - A\mathbf{z}[k-1] - B\mathbf{u}[k-1] - \mathbf{w}[k],$$

so that if the sensors truthfully report $\mathbf{z}[t] \equiv \mathbf{y}[t]$, then, $\mathbf{v} \equiv 0$. As will be shown, the sequence $\{\mathbf{v}\}$ has the interpretation of the additive noise distortion introduced by the malicious sensors to the process noise. A similar definition can also be provided for a partially observed system, as we will show in Section 6.7. In this case, the sequence $\{\mathbf{v}\}$ has the interpretation of the additive distortion introduced by the malicious sensors to the innovations process. Now, based on $\{\mathbf{v}\}$, we define the following quantity.

Definition 7. *We call*

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} \|\mathbf{v}[k]\|^2$$

as the additive noise distortion power *of the malicious sensors.*

The "fundamental security guarantee," described in the sequel, provided by Dynamic Watermarking is that the additive noise distortion power is restricted to be zero if the malicious sensors are to remain undetected. We establish this in several linear control system contexts in this chapter.

6.4 Active Defense for Networked Cyber-Physical Systems: The SISO Case with Gaussian noise

In this section, to illustrate the results in a simple context, we focus on single-input, single-output linear stochastic dynamical systems with Gaussian noise. The system is described by

$$x[t+1] = ax[t] + bu[t] + w[t+1], \tag{6.6}$$

where $a, b, x[t], u[t], w[t] \in \mathbb{R}$, with $\{w[t]\}$ being zero-mean i.i.d. Gaussian process noise of variance σ_w^2. The actuator wishes to implement a control law $\{g_t\}$, i.e., it wishes to implement $u[t] = g_t(x^t)$, where $x^t \coloneqq (x[0], x[1], \cdots, x[t])$. However, the actuator does not have access to x^t. It relies on a sensor that measures $x[t]$. However, since the sensor could be malicious, it reports

measurements $z[t]$ to the actuator, where $z[t]$ could differ from $x[t]$. We consider an *honest* actuator that is meant to, and implements the control policy $\{g\}$, but adds a private excitation $\{e\}$ as a defense. Specifically, the actuator applies to the system the input

$$u[t] = g_t(z^t) + e[t]. \tag{6.7}$$

Note that even though it implements the control policy $\{g\}$, the policy is applied to the measurements $z[t]$ reported by the sensor, which could differ from the true output $x[t]$. The private excitation $e[\cdot]$ added is independent and identically distributed (i.i.d.) and Gaussian of mean 0 and variance σ_e^2. Therefore, the system evolves in closed-loop as

$$x[t+1] = ax[t] + bg_t(z^t) + be[t] + w[t+1]. \tag{6.8}$$

We propose that the honest actuator perform certain "tests" to check if the sensor is malicious or not. Towards developing these tests, note that the actual sequence of states $\{x[t]\}$ of the system satisfies

$$x[t+1] - ax[t] - bg_t(z^t) = be[t] + w[t+1]. \tag{6.9}$$

Therefore, we have

$$\{x[t+1] - ax[t] - bg_t(z^t)\}_t \sim i.i.d.\ \mathcal{N}(0, b^2\sigma_e^2 + \sigma_w^2), \tag{6.10}$$

and

$$\{x[t+1] - ax[t] - bg_t(z^t) - be[t]\}_t \sim i.i.d.\ \mathcal{N}(0, \sigma_w^2). \tag{6.11}$$

Based on the above observations, we propose that the actuator perform the following natural tests for variance to check if the sensor is honestly reporting $x[t]$. The actuator checks if the reported

sequence $\{z[t]\}$ satisfies conditions (6.10) and (6.11), which the true output $\{x[t]\}$ would satisfy if the sensor were truthfully reporting $z[t] \equiv x[t]$. We write the tests in an asymptotic form below, as a test conducted over an infinite time interval. They can be reduced to statistical tests over a finite time interval in standard ways, which we elaborate on in Section 6.11.

1) **Actuator Test 1:** Check if the reported sequence of measurements $\{z[t]\}$ satisfies

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (z[k+1] - az[k] - bg_k(z^k) - be[k])^2 = \sigma_w^2. \tag{6.12}$$

2) **Actuator Test 2:** Check if the reported sequence of measurements $\{z[t]\}$ satisfies

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (z[k+1] - az[k] - bg_k(z^k))^2 = (b^2\sigma_e^2 + \sigma_w^2). \tag{6.13}$$

Define

$$v[t+1] \coloneqq z[t+1] - az[t] - bg_t(z^t) - be[t] - w[t+1],$$

so that for an honest sensor which reports $z[t] \equiv x[t]$, $v[t] = 0 \ \ \forall t$. We term the quantity

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} v^2[k]$$

the *additive noise distortion power* of a malicious sensor for reasons explained later. The ensuing theorem proves that a malicious sensor with only zero additive noise distortion power can pass the above two tests, thereby remaining undetected.

Theorem 5. *If $\{z[t]\}$ passes tests (6.12) and (6.13), thereby remaining undetected, then,*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} v^2[k] = 0. \tag{6.14}$$

Proof. Since $\{z\}$ satisfies (6.12), we have,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} (v[k] + w[k])^2 = \sigma_w^2. \tag{6.15}$$

Hence,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} v^2[k] + 2v[k]w[k] + w^2[k] = \sigma_w^2.$$

Since $\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} w^2[k] = E\{w^2[k]\} = \sigma_w^2$, we have

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} v^2[k] + \lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} 2v[k]w[k] = 0. \tag{6.16}$$

Since $\{z\}$ also satisfies (6.13), we have,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} (v[k] + be[k-1] + w[k])^2 = b^2\sigma_e^2 + \sigma_w^2.$$

So,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} (v[k] + w[k])^2 + \lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} b^2e^2[k-1] + \lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} 2be[k-1](v[k] + w[k])$$
$$= b^2\sigma_e^2 + \sigma_w^2.$$

Using (6.15), and the fact that $\{e\}$ has variance σ_e^2, the above reduces to

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} e[k-1](v[k] + w[k]) = 0.$$

Invoking the fact that $e[k-1]$ and $w[k]$ are independent $\forall k$ and zero mean, it further reduces to

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} e[k-1]v[k] = 0. \tag{6.17}$$

The above equation implies that the sequence $\{v\}$, added by the sensor, must be empirically uncorrelated with the actuator's private noise sequence $\{e\}$.

Let $\mathcal{S}_k \coloneqq \sigma(x^k, z^k, e^{k-2})$, $\widehat{w}[k] \coloneqq E[w[k]\big|\mathcal{S}_k]$. Since

$$w[k] = x[k] - ax[k-1] - bg_{k-1}(z^{k-1}) - be[k-1],$$

we have

$$(x^{k-2}, e^{k-2}) \to (x[k-1], x[k], z^k) \to w[k]$$

forming a Markov chain. Consequently, $\widehat{w}[k] \coloneqq E[w[k]\big|\sigma(e^{k-2}, x^{k-2}, x[k-1], x[k], z^k)] = E[w[k]\big|\sigma(x[k-1], x[k], z^k)]$. Since $x[k]-ax[k-1]-bg_{k-1}(z^{k-1})$ (which is equal to $be[k-1]+w[k]$) is i.i.d. Gaussian for different k, we have [54]

$$\widehat{w}[k] = \frac{\sigma_w^2}{b^2\sigma_e^2 + \sigma_w^2}(be[k-1] + w[k]) = \beta(be[k-1] + w[k]), \tag{6.18}$$

where $\beta \coloneqq \frac{\sigma_w^2}{b^2\sigma_e^2+\sigma_w^2} < 1$.

Let $\widetilde{w}[k] \coloneqq w[k] - \widehat{w}[k]$. Then, $(\widetilde{w}[k-1], \mathcal{S}_k)$ is a Martingale difference sequence. This is because $\widetilde{w}[k-1] \in \mathcal{S}_k$, and

$$E[\widetilde{w}[k]\ \big|\mathcal{S}_k] = 0. \tag{6.19}$$

We also have $v[k] \in \mathcal{S}_k$ (in fact, $v[k] \in \sigma(x^k, z^k)$). Hence, Martingale Stability Theorem (MST) [49] applies, and we have

$$\sum_{k=1}^{T} v[k]\widetilde{w}[k] = o(\sum_{k=1}^{T} v^2[k]) + O(1). \tag{6.20}$$

Now,

$$\sum_{k=1}^{T} v[k]w[k] = \sum_{k=1}^{T} v[k](\widehat{w}[k] + \widetilde{w}[k]) = \sum_{k=1}^{T} v[k]\widehat{w}[k] + o(\sum_{k=1}^{T} v^2[k]) + O(1).$$

Employing the specific form of the estimate (6.18), we have from the above,

$$\sum_{k=1}^{T} v[k]w[k] = \beta b \sum_{k=1}^{T} v[k]e[k-1] + \beta \sum_{k=1}^{T} v[k]w[k] + o(\sum_{k=1}^{T} v^2[k]) + O(1).$$

Hence,

$$\sum_{k=1}^{T} v[k]w[k] = \frac{\beta b}{1-\beta} \sum_{k=1}^{T} v[k]e[k-1] + o(\sum_{k=1}^{T} v^2[k]) + O(1).$$

From (6.17), we have $\sum_{k=1}^{T} v[k]e[k-1] = o(T)$. It follows that

$$\sum_{k=1}^{T} v[k]w[k] = o(\sum_{k=1}^{T} v^2[k]) + o(T) + O(1). \tag{6.21}$$

So,

$$\sum_{k=1}^{T} v^2[k] + \sum_{k=1}^{T} 2v[k]w[k] = (1 + o(1))(\sum_{k=1}^{T} v^2[k]) + o(T) + O(1)$$

Dividing the above equation by T, taking the limit as $T \to \infty$, and invoking (6.16) completes the proof. □

Remark: Note that the only sources of uncertainty in the system are the initial state of the system $x[0]$ and the sequence of noise realizations $\{w[1], w[2], w[3], \cdots\}$. The sensor reporting a sequence of measurements is equivalent to it reporting a sequence of process noise realizations, since the actuator expects $z[t+1] - az[t] - bg_t(z^t) - be[t]$, which it can compute, to be equal to the process noise $w[t+1]$. From the definition of $v[t]$, we have

$$z[t+1] - az[t] - bg_t(z^t) - be[t] = w[t+1] + v[t+1].$$

The left hand side of the above equation can be computed by the actuator, and therefore, it can also compute the sequence $\{w+v\}$. What the theorem states is that a malicious sensor cannot distort the

noise realization $\{w[1], w[2], w[3], \cdots\}$ beyond adding a zero-power sequence to it. Suppose the control law g has been designed to provide good noise regulation performance of a stable system, then the performance of the system is good, subject only to the slightly increased cost of the private excitation of variance σ_e^2. As we discuss in Section 6.11, this can be reduced to a low enough value that permits detection of a malicious sensor within a specified delay with an acceptable level of false alarm probability.

Theorem 6. *Suppose $|a| < 1$, i.e., the system is stable.*

(i) Define the distortion $d[t] := z[t] - x[t]$. Then,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} d^2[k] = 0.$$

(ii) If the malicious sensor is to remain undetected, the mean-square performance of $x[t]$ is the same as the reported mean-square performance $z[t]$ that the actuator believes it is:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} x^2[k] = \lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} z^2[k].$$

(iii) Suppose the control law is $u^g(t) = fx(t)$ with $|a + bf| < 1$. The malicious sensor cannot compromise the performance of the system if it is to remain undetected, i.e., the mean-square performance of the system is

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} x^2[k] = \frac{\sigma_w^2 + b^2\sigma_e^2}{1 - |a + bf|^2}.$$

Proof. Note that

$$\begin{aligned} d[k+1] &= z[k+1] - x[k+1] \\ &= (az[k] + bg_k(z^k) + be[k] + w[k+1] + v[k+1]) - (ax[k] + bg_k(z^k) + be[k] + w[k+1]) \\ &= a(z[k] - x[k]) + v[k+1] = ad[k] + v[k+1]. \end{aligned}$$

The distortion experienced by the actuator can therefore be thought of as the output of a linear dynamical system driven by an input sequence $\{v[t]\}$ satisfying (6.14). Therefore,

$$d[k] = \sum_{n=0}^{k-1} a^n v[k-n],$$

where $\lim_{T\to\infty} \frac{1}{T} \sum_{p=1}^{T} v^2[p] = 0$. From the stability of a, it follows that

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} d^2[k] = 0. \tag{6.22}$$

Note that $x[k] = z[k] - d[k]$. Hence $x^2[k] = z^2[k] + d^2[k] - 2(\gamma z[k])(\gamma^{-1} d[k])$. Now $|2(\gamma z[k])(\gamma^{-1} d[k])| \leq (\gamma^2 z^2[k]) + (\gamma^{-2} d^2[k])$. Therefore,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} x^2[k] \leq \lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (1+\gamma^2) z^2[k] + (1+\gamma^{-2}) d^2[k].$$

Since the result is true for any $\gamma > 0$, taking the limit $\gamma \to 0$ and noting that the mean-square of d is 0 gives

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} x^2[k] \leq \lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} z^2[k]. \tag{6.23}$$

Similarly, since $z[k] = x[k] + d[k]$, we have $z^2[k] = x^2[k] + d^2[k] + 2(\gamma x[k])(\gamma^{-1} z[k])$. Repeating the same argument as before, we have

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} z^2[k] \leq \lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (1+\gamma^2) x^2[k] + (1+\gamma^{-2}) d^2[k].$$

Taking the limit as $\gamma \to 0$, and noting that mean-square of d is 0, the above reduces to

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} z^2[k] \leq \lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} x^2[k]. \tag{6.24}$$

The second result follows from (6.23) and (6.24). The third result is immediate noting that the mean-square of $\{z\}$ converges to $\frac{\sigma_w^2+b^2\sigma_e^2}{1-|a+bf|^2}$. □

6.5 Active Defense for Networked Cyber-Physical Systems: The SISO ARX Case

The results developed in the previous section can be extended to a more general ARX system model. Specifically, we consider a unit delay, strictly minimum phase, single-input, single-output system described by

$$y[t+1] = -\sum_{m=0}^{p} a_m y[t-m] + \sum_{r=0}^{h} b_r u[t-r] + w[t+1], \tag{6.25}$$

where $a_m, b_r, y[t], u[t], w[t] \in \mathbb{R}$, $b_0 \neq 0$, and with $\{w[t]\}$ being zero-mean i.i.d. Gaussian process noise of variance σ_w^2. Let q^{-1} denote the backward shift operator. The above system can equivalently be expressed as

$$A(q^{-1})y[t] = q^{-1}B(q^{-1})u[t] + w[t], \tag{6.26}$$

where $A(q^{-1}) \coloneqq 1 + a_0q^{-1} + a_1q^{-2} + ... + a_pq^{-(p+1)}$, and $B(q^{-1}) \coloneqq b_0 + b_1q^{-1} + ... + b_hq^{-h}$, with $B(q^{-1})$ being strictly minimum phase, i.e., all its roots lie strictly outside the unit circle.

We consider an honest actuator that is meant to, and implements the control policy $\{g\}$, and adds a private excitation $\{e\}$ as a defense. Unlike in the system considered in Section-6.4, the output of the ARX system at any particular time depends on the past inputs. Specifically, the output at any time instant contains contributions of private excitation injected in the past. Hence, simply injecting an i.i.d. sequence of Gaussian random variables will not result in an output distribution that is i.i.d. across time.

However, since the actuator knows the past values of the private excitation, and also the transfer function of the system, it can perform "pre-equalization" by filtering the private excitation sequence before injecting it into the system. The filter, which we refer to as the pre-equalizer, has to be chosen in such a way that the component of the private excitation that appears in the output of the

plant is an i.i.d. sequence of mean 0 and variance $b_0^2\sigma_e^2$. We let

$$e'[t] \coloneqq -\frac{1}{b_0}(b_1 e'[t-1] + b_2 e'[t-2] + \ldots + b_h e'[t-h]) + e[t],$$

where $e[t]$ is i.i.d. and Gaussian of mean 0 and variance σ_e^2. Since the system is strictly minimum phase, this is a stable generation of an excitation sequence [54].

With a pre-equalizer in place, the effective input applied by the actuator is

$$u[t] = g_t(z^t) + e'[t]. \tag{6.27}$$

where $\{e'\}_t$ is the output of the pre-equalizer, and z^t is the sequence of measurements reported by the sensor up to time t. Therefore, the system evolves in closed-loop as

$$y[t+1] = -\sum_{m=0}^{p} a_m y[t-m] + \sum_{r=0}^{h} b_r g_{t-r}(z^{t-r}) + b_0 e[t] + w[t+1], \tag{6.28}$$

where $\{e[t]\}$ is a sequence of i.i.d. Gaussian random variables with mean 0 and variance σ_e^2. Hence, from the point of view of actuator tests and associated results, the above problem is similar to that described in Section-6.4.

We propose that the actuator perform the following tests to check if the sensor is reporting the measurements honestly.

1) **Actuator Test 1:** Check if the reported sequence of measurements $\{z[t]\}$ satisfies

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (z[k+1] + \sum_{m=0}^{p} a_m z[k-m] - \sum_{r=0}^{h} b_r g_{k-r}(z^{k-r}) - b_0 e[k])^2 = \sigma_w^2. \tag{6.29}$$

2) **Actuator Test 2:** Check if the reported sequence of measurements $\{z[t]\}$ satisfies

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (z[k+1] + \sum_{m=0}^{p} a_m z[k-m] - \sum_{r=0}^{h} b_r g_{k-r}(z^{k-r}))^2 = (b_0^2\sigma_e^2 + \sigma_w^2). \tag{6.30}$$

Consequently, Theorem 1 and Theorem 2 can be extended as follows.

Theorem 7. *1) Let* $v[t+1] \coloneqq z[t+1] + \sum_{m=0}^{p} a_m z[k-m] - \sum_{r=0}^{h} b_r g_{k-r}(z^{k-r}) - b_0 e[t] - w[t+1]$*, so that for an honest sensor which reports* $z[t] \equiv y[t]$*,* $v[t] = 0 \ \ \forall t$*. If* $\{z[t]\}$ *passes tests (6.29) and (6.30), then,*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} v^2[k] = 0. \tag{6.31}$$

2) If the malicious sensor is to remain undetected, the mean-square performance of $\{y\}$ *is what the actuator believes it is, which is the mean-square performance of* $\{z\}$*. Hence if the control law was designed to provide a certain mean-square performance, then that is the value that is indeed attained.*

Proof. Omitted, since the proof follows the same sequence of arguments as the proof of Theorem 1. □

In the following section, we extend the technique to systems with arbitrary delay and colored noise, i.e., the more general ARMAX model, see [55].

6.6 Active Defense for Networked Cyber-Physical Systems: The SISO ARMAX Case

A general ARMAX system with arbitrary but finite delay is a model that is encountered often in process control. In this section, we develop a Dynamic Watermarking scheme to secure such systems. We show that by commanding the actuator to inject private excitation whose spectrum is matched to that of the colored process noise entering the system, it can be ensured that a malicious sensor is constrained to distorting the process noise, the only quantity unknown to the actuator, by at most a zero-power signal. This amounts to a form of Internal Model Principle [56] for Dynamic Watermarking, and is a phenomenon that doesn't emerge in the analysis of a simple SISO or an ARX system treated in the previous sections.

A general ARMAX system with finite delay l is described by

$$y[t] = -\sum_{k=1}^{p} a_k y[t-k] + \sum_{k=0}^{h} b_k u[t-l-k] + \sum_{k=0}^{r} c_k w[t-k]. \tag{6.32}$$

Without loss of generality, we assume that $c_0 = 1$. Let q^{-1} be the backward shift operator, so that $q^{-1}y[t] = y[t-1]$. Then, the above system can be expressed as

$$A(q^{-1})y[t] = q^{-l}B(q^{-1})u[t] + C(q^{-1})w[t], \tag{6.33}$$

where $a_0 \coloneqq 1$, $A(q^{-1}) \coloneqq a_0 + a_1q^{-1} + ... + a_pq^{-p}$, $B(q^{-1}) \coloneqq b_0 + b_1q^{-1} + ... + b_hq^{-h}$, $C(q^{-1}) \coloneqq c_0 + c_1q^{-1} + ... + c_rq^{-r}$, and $B(q^{-1})$ and $C(q^{-1})$ are assumed to be strictly minimum phase. To secure the system, the actuator applies the control

$$u[t] = u^g[t] + B^{-1}(q^{-1})C(q^{-1})e[t], \tag{6.34}$$

where $u^g[t]$ is the control policy-specified input sequence and $\{e\}$ is a sequence of i.i.d Gaussian random variables of zero mean and variance σ_e^2, denoting the actuator node's private excitation. Since $B(q^{-1})$ is assumed to be minimum phase, (6.34) is a stable generation of the control input $u[t]$. Consequently, the output of the system obeys

$$A(q^{-1})y[t] = q^{-l}B(q^{-1})u^g[t] + q^{-l}C(q^{-1})e[t] + C(q^{-1})w[t], \tag{6.35}$$

implying that

$$y[t] = -\sum_{k=1}^{p} a_k y[t-k] + \sum_{k=0}^{h} b_k u^g[t-l-k] + \sum_{k=0}^{r} c_k e[t-l-k] + \sum_{k=0}^{r} c_k w[t-k]. \tag{6.36}$$

While the $q-$domain derivation is just formal, the above difference equation can be obtained rigorously.

Define $\lambda[t] \coloneqq e[t-l] + w[t]$. Then, the output of the system can be expressed as

$$y[t] = -\sum_{k=1}^{p} a_k y[t-k] + \sum_{k=0}^{h} b_k u^g[t-l-k] + \sum_{k=0}^{r} c_k \lambda[t-k]. \tag{6.37}$$

We now develop tests that the actuator should perform to detect maliciousness of the sensor. The

fundamental idea behind the tests is to check if the prediction-error of the reported sequence $\{z\}$ has the appropriate statistics. Specifically, the actuator processes the reported measurements with a prediction-error filter defined through the following recursion for all $t \geq 0$.

$$z_{t|t-1} = -\sum_{k=1}^{p} a_k z[t-k] + \sum_{k=0}^{h} b_k u^g[t-l-k] + \sum_{k=1}^{r} c_k \widetilde{z}[t-k], \tag{6.38}$$

$$\widetilde{z}[t] = z[t] - z_{t|t-1}. \tag{6.39}$$

The filter is assumed to be initialized with $\widetilde{z}[-k] = \lambda[-k], k \in \{-1, -2, ..., -r\}$. It is easy to verify that the above filter produces $\widetilde{z}[t] \equiv \lambda[t]$ if $z[t] \equiv y[t]$. Based on this observation, we propose that the actuator perform the following tests.

1. **Actuator Test 1:** The actuator checks if

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (\widetilde{z}[k] - e[k-l])^2 = \sigma_w^2. \tag{6.40}$$

2. **Actuator Test 2:** The actuator checks if

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} \widetilde{z}^2[k] = \sigma_w^2 + \sigma_e^2. \tag{6.41}$$

The following theorem shows that the above tests suffice to ensure that a malicious sensor cannot introduce any distortion beyond addition of a zero-power signal to the process noise. i.e., a malicious sensor of only zero additive noise distortion power can pass the above tests to remain undetected.

Theorem 8. *Define* $v[t] \coloneqq \sum_{k=0}^{p} a_k z[t-k] - \sum_{k=0}^{h} b_k u^g[t-l-k] - \sum_{k=0}^{r} c_k \lambda[t-k]$*, so that for an honest sensor reporting* $z \equiv y$*,* $v \equiv 0$*. If the sensor passes tests (6.40) and (6.41), then,*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} v^2[k] = 0.$$

Proof. From (6.38) and (6.39), one gets

$$\sum_{k=0}^{r} c_k \widetilde{z}[t-k] = z[t] + \sum_{k=1}^{p} a_k z[t-k] - \sum_{k=0}^{h} b_k u^g[t-l-k] = \sum_{k=0}^{r} c_k \lambda[t-k] + v[t]. \quad (6.42)$$

Equivalently, assuming appropriate initial conditions, one has

$$C(q^{-1})\widetilde{z}[t] = C(q^{-1})\lambda[t] + v[t].$$

This gives

$$\widetilde{z}[t] = \lambda[t] + v_f[t],$$

where $v_f[t] \coloneqq C^{-1}(q^{-1})v[t]$. Now,

$$\widetilde{z}[t] = \lambda[t] + v_f[t] = w[t] + e[t-l] + v_f[t]. \quad (6.43)$$

Since $\{\widetilde{z}\}$ passes (6.40), we have from the above,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (v_f[k] + w[k])^2 = \sigma_w^2. \quad (6.44)$$

This gives

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} v_f^2[k] + 2v_f[k]w[k] = 0. \quad (6.45)$$

Since $\{\widetilde{z}\}$ also passes (6.41), we have

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (v_f[k] + w[k] + e[k-l])^2 = \sigma_w^2 + \sigma_e^2. \quad (6.46)$$

This gives

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1} v_f^2[k] + 2v_f[k]w[k] + 2v_f[k]e[k-l] = 0.$$

Combining the above with (6.45), we have

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1} v_f[k]e[k-l] = 0. \tag{6.47}$$

Comparing (6.45) and (6.47) with (6.16) and (6.17), we see that the filtered distortion measure $\{v_f\}$ behaves the same way $\{v\}$ does in the white noise case. Proceeding the same way as in the proof of Theorem 1, one arrives at

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1} v_f^2[k] = 0. \tag{6.48}$$

Since $v[k] = C(q^{-1})v_f[k]$, and $C(q^{-1})$ is minimum phase, it follows that $\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1} v^2[k] = 0$. □

6.7 Active Defense for Networked Cyber-Physical Systems: SISO Systems with Partial Observations

In this section, we address SISO systems with noisy, partial observations. We consider a p^{th} order single input single output system described by

$$\mathbf{x}[t+1] = A\mathbf{x}[t] + Bu[t] + \mathbf{w}[t+1], \tag{6.49}$$

$$y[t+1] = C\mathbf{x}[t+1] + n[t+1]. \tag{6.50}$$

where $\mathbf{x}[t] \in \mathbb{R}^p$, and A, B, and C are known matrices of appropriate dimensions. The actuator, being honest, applies the input

$$u[t] = g_t(z^t) + e[t], \tag{6.51}$$

where $g_t(z^t)$ is the control policy-specified input, and $e[t] \sim \mathcal{N}(0, \sigma_e^2)$ is a sequence of i.i.d random variables denoting the actuator node's private excitation. Consequently, the system evolves as

$$\mathbf{x}[t+1] = A\mathbf{x}[t] + Bg_t(z^t) + be[t] + \mathbf{w}[t+1], \tag{6.52}$$

$$y[t+1] = C\mathbf{x}[t+1] + n[t+1]. \tag{6.53}$$

Let $z[t]$ be the measurement reported by the sensor at time t. Since the sensor can be malicious, $z[t]$ need not equal $y[t]$ for every t.

The actuator performs Kalman filtering on the reported measurements $\{z\}$ as follows.

$$\widehat{\mathbf{x}}_F(k+1|k) = A\widehat{\mathbf{x}}_F(k|k) + Bg_k(z^k) + Be[k], \tag{6.54}$$

$$\widehat{\mathbf{x}}_F(k+1|k+1) = A\widehat{\mathbf{x}}_F(k|k) + Bg_k(z^k) + Be[k] + K_k\nu_F[k+1], \tag{6.55}$$

where $\nu_F[k+1] \coloneqq z[k+1] - C\widehat{\mathbf{x}}_F(k+1|k)$, denotes the (possibly) faulty innovations computed by the actuator, and K_k is the Kalman gain at time k.

We also define a Kalman filter that operates on the true measurements $y[t]$ as follows.

$$\widehat{\mathbf{x}}_R(k+1|k) = A\widehat{\mathbf{x}}_R(k|k) + Bg_k(z^k) + Be[k], \tag{6.56}$$

$$\widehat{\mathbf{x}}_R(k+1|k+1) = A\widehat{\mathbf{x}}_R(k|k) + Bg_k(z^k) + Be[k] + K_k\nu_R[k+1], \tag{6.57}$$

where $\nu_R[k+1] \coloneqq y[k+1] - C\widehat{\mathbf{x}}_R(k+1|k)$ is the "real innovations" or "true innovations" of the system. Of course, the actuator cannot implement the latter Kalman filter since it may not receive $y[k]$ from the sensor.

We now define

$$\mathbf{v}[k+1] \coloneqq \widehat{\mathbf{x}}_F(k+1|k+1) - A\widehat{\mathbf{x}}_F(k|k) - Bg_k(z^k) - Be[k] - K_k\nu_R[k+1], \tag{6.58}$$

so that for an honest sensor which reports $z \equiv y$, we have $\nu \equiv \nu_R$ and consequently, $\mathbf{v} \equiv 0$. The actuator performs the following two tests to detect maliciousness of the sensor.

1) **Actuator Test 1:** Actuator checks if

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} e[k](\widehat{\mathbf{x}}_F(k+1|k+1) - A\widehat{\mathbf{x}}_F(k|k) - Bg_k(z^k) - Be[k]) = 0 \tag{6.59}$$

2) **Actuator Test 2:** Actuator checks if

$$\begin{aligned}\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} &(\widehat{\mathbf{x}}_F(k+1|k+1) - A\widehat{\mathbf{x}}_F(k|k) - Bg_k(z^k) - Be[k]) \\ &(\widehat{\mathbf{x}}_F(k+1|k+1) - A\widehat{\mathbf{x}}_F(k|k) - Bg_k(z^k) - Be[k])^T = \sigma_R^2 K K^T,\end{aligned} \tag{6.60}$$

where σ_R^2 denotes the variance of the true innovations process, and K is the steady-state Kalman gain. We assume that (A, C) is observable so that the algebraic Riccati equation associated with the Kalman gain has a unique nonnegative definite solution [54]. The following theorem shows that the above tests suffice to ensure that a malicious sensor of only zero additive noise distortion power can pass the tests to remain undetected.

Theorem 9. *Suppose that the reported sequence of measurements passes the tests (6.59) and (6.60). Then,*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} \|\mathbf{v}[k+1]\|^2 = 0. \tag{6.61}$$

Proof. Since the reported sequence of measurements passes (6.59), we have

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}e[k](K_k\nu_R[k+1]+\mathbf{v}[k+1])=0.$$

Since the true innovations are independent of the zero-mean private excitation, it follows that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}e[k]\mathbf{v}[k+1]=0. \tag{6.62}$$

Since the reported sequence of measurements also passes (6.60), we have

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}(K_k\nu_R[k+1]+\mathbf{v}[k+1])(K_k\nu_R[k+1]+\mathbf{v}[k+1])^T=\sigma_R^2KK^T.$$

Simplifying, the above gives

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}(K_k\nu_R[k+1]\mathbf{v}^T[k+1])+(K_k\nu_R[k+1]\mathbf{v}^T[k+1])^T+(\mathbf{v}[k+1]\mathbf{v}^T[k+1])=0. \tag{6.63}$$

We define $\mathcal{S}_k\coloneqq\sigma(\widehat{\mathbf{x}}_R^{k|k},\widehat{\mathbf{x}}_F^{k|k},z^k,e^{k-2},y^{k-1})$, where $\widehat{\mathbf{x}}_R^{k|k}\coloneqq\{\widehat{\mathbf{x}}_R(k|k),\widehat{\mathbf{x}}_R(k-1|k-1),...,\widehat{\mathbf{x}}_R(-1|-1)\}$, and $\widehat{\mathbf{x}}_F^{k|k}$ is defined likewise. Define $\widehat{\nu}_R[k]\coloneqq E[\nu_R[k]\big|\mathcal{S}_k]$. From the Kalman filtering equations, we have

$$K_{k-1}\nu_R[k]=\widehat{\mathbf{x}}_R(k|k)-A\widehat{\mathbf{x}}_R(k-1|k-1)-Bg_{k-1}(z^{k-1})-Be[k-1]$$

implying that

$$(\widehat{\mathbf{x}}_R^{k-2|k-2},\widehat{\mathbf{x}}_F^{k|k},e^{k-2},y^{k-1})\to(\widehat{\mathbf{x}}_R(k-1|k-1),\widehat{\mathbf{x}}_R(k|k),z^k)\to\nu_R[k]$$

forms a Markov chain. Therefore, $\widehat{\nu}_R[k]\coloneqq E[\nu_R[k]\big|\sigma(\widehat{\mathbf{x}}_R^{k|k},\widehat{\mathbf{x}}_F^{k|k},z^k,e^{k-2},y^{k-1})]=E[\nu_R[k]\big|\sigma(\widehat{\mathbf{x}}_R(k-$

$1|k-1), \widehat{\mathbf{x}}_R(k|k), z^k)]$. Therefore, we have

$$K_{k-1}\widehat{\nu}_R[k] = K_\nu(K_{k-1}\nu_R[k] + Be[k-1]), \tag{6.64}$$

where $K_\nu := \sigma_R^2 K_{k-1}K_{k-1}^T(\sigma_R^2 K_{k-1}K_{k-1}^T + \sigma_e^2 BB^T)^{-1}$. Define $K_{k-1}\widetilde{\nu}_R[k] := K_{k-1}\nu_R[k] - K_{k-1}\widehat{\nu}_R[k]$. Then, $(K_{k-1}\widetilde{\nu}_R[k-1], \mathcal{S}_k)$ is a martingale difference sequence. This is because, $\widetilde{\nu}_R[k-1] \in \mathcal{S}_k$, and

$$E[\widetilde{\nu}_R[k]\big|\mathcal{S}_k] = 0.$$

Also, $\mathbf{v}[k] \in \mathcal{S}_k$. Hence, MST applies, and we have

$$\sum_{k=1}^{T} K_{k-1}\widetilde{\nu}_R[k]\mathbf{v}^T[k] = \begin{bmatrix} o(\sum_{k=0}^{T-1} v_1^2[k+1]) & \dots & o(\sum_{k=0}^{T-1} v_p^2[k+1]) \\ o(\sum_{k=0}^{T-1} v_1^2[k+1]) & \dots & o(\sum_{k=0}^{T-1} v_p^2[k+1]) \\ \vdots & \dots & \vdots \\ o(\sum_{k=0}^{T-1} v_1^2[k+1]) & \dots & o(\sum_{k=0}^{T-1} v_p^2[k+1]) \end{bmatrix} + O(1), \tag{6.65}$$

where $v_i[k]$ denotes the i^{th} component of $\mathbf{v}[k]$. Now, since $K_k\nu_R[k] = K_k\widehat{\nu}_R[k] + K_k\widetilde{\nu}_R[k]$, from (6.64), we have

$$K_{k-1}\nu_R[k] = (I - K_\nu)^{-1}Be[k-1] + (I - K_\nu)^{-1}K_{k-1}\widetilde{\nu}_R[k]. \tag{6.66}$$

Substituting this in (6.63) and equating the diagonals using (6.62) and (6.65) completes the proof.

□

6.8 Active Defense for Networked Cyber-Physical Systems: MIMO Systems with Gaussian Noise

In this section, we investigate multiple-input-multiple-output (MIMO) linear stochastic dynamical systems. They pose additional challenges as compared to SISO systems since malicious sensors in a MIMO system can attempt to collude so as to prevent the other nodes from detecting

their presence. In this section, we show how the actuators can prevent the malicious nodes from introducing "excessive" distortion, lest they expose their presence.

An m input MIMO Linear Dynamical System of order n is described by

$$\mathbf{x}[t+1] = A\mathbf{x}[t] + B\mathbf{u}[t] + \mathbf{w}[t+1] \tag{6.67}$$

where $\mathbf{x}[t] \in \mathbb{R}^n$, $\mathbf{u}[t] \in \mathbb{R}^m$, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, and $\{\mathbf{w}\}$ is a zero-mean i.i.d. sequence of Gaussian random vectors with covariance matrix $\sigma_w^2 I_n$. We consider the case where $\mathbf{x}$ is perfectly observed.

Let $\mathbf{z}[\mathbf{t}]$ be the measurements reported by the sensors to the other nodes in the system. Note that since the nodes in the system are allowed to exchange reported measurements among them, all malicious sensors have to be consistent and report the same (but possibly erroneous) measurements to all honest nodes in the system. A history-dependent control policy is assumed to be in place, which dictates that the i^{th} input to the system at time t be

$$u_i^g[t] = g_t^i(\mathbf{z}^t). \tag{6.68}$$

As reasoned before, to secure the system, each actuator superimposes on the input specified by the control policy, an additional zero-mean private excitation that it draws from a distribution, here Gaussian, that is made public. It should be noted that while the *distribution* is made public, the actual values of the excitation are not revealed by the actuator to any other node. The private excitation value drawn at each time t is chosen to be independent of its private excitation values at all the other time instants, of the private excitation values of other actuator nodes, and of the control policy-specified input. Therefore, actuator i applies at time t the input

$$u_i[t] = u_i^g[t] + e_i[t] = g_t^i(\mathbf{z}^t) + e_i[t], \tag{6.69}$$

where $e_i[t] \sim \mathcal{N}(0, \sigma_e^2)$ is independent of $e_j[k]$ for $(j,k) \neq (i,t)$, $\mathbf{x}[m], \mathbf{z}[m]$ for $m \leq t$, and

$\mathbf{w}[n]$ for all n. We assume that all actuators are honest, meaning that they apply control inputs in accordance to (6.69).

We propose the following tests to be performed by each actuator i.

1) **Actuator Test 1:** Actuator i checks if the reported sequence of measurements $\{\mathbf{z}\}$ satisfies

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}(\mathbf{z}[k+1]-A\mathbf{z}[k]-Bg_k(\mathbf{z}^k))(\mathbf{z}[k+1]-A\mathbf{z}[k]-Bg_k(\mathbf{z}^k))^T = \sigma_e^2 BB^T+\sigma_w^2 I_n \tag{6.70}$$

2) **Actuator Test 2:** Actuator i checks if the reported sequence of measurements $\{\mathbf{z}\}$ satisfies

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}e_i[k](\mathbf{z}[k+1]-A\mathbf{z}[k]-Bg_k(\mathbf{z}^k)) = B_{\cdot,i}\sigma_e^2 \tag{6.71}$$

In Actuator Test 2, we have simplified the test to directly check the quantity that is important, which in this is case is the analog of (6.12). We note that for any set of conditions to be an admissible test, they have to be (i) checkable by the sensors and actuators based on their observations and what is reported to them, and (ii) satisfied if all parties are honest. The above two conditions satisfy these two conditions.

Define

$$\mathbf{v}[t+1] \coloneqq \mathbf{z}[t+1]-A\mathbf{z}[t]-Bg_t(\mathbf{z}^t)-B\mathbf{e}[t]-\mathbf{w}[t+1],$$

so that if $\mathbf{z}[t]\equiv\mathbf{x}[t]$, $\mathbf{v}[t]\equiv 0$. It is easy to see that as in the case of SISO systems, the sequence $\{\mathbf{v}\}$ has the intepretation as the additive noise distortion of the process noise deliberately introduced by the malicious sensors. We term the quantity

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}||\mathbf{v}[k]||^2$$

the *additive noise distortion power* of the malicious sensors. The following theorem, akin to

Theorem 1, proves that malicious sensors of only zero additive noise distortion power can pass the above tests, thereby remaining undetected.

Theorem 10. *Suppose that the reported sequence of measurements passes the tests (6.70), and (6.71). If B is of rank n, then,*

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}||\mathbf{v}[k]||^2=0. \tag{6.72}$$

Proof. Since $\{\mathbf{z}\}$ satisfies (6.71), we have, $\forall i\in\{1,2,...,m\}$,

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}e_i[k](B\mathbf{e}[k]+\mathbf{w}[k+1]+\mathbf{v}[k+1])=\sigma_e^2B_{\cdot(i)}.$$

It follows that for all $i\in\{1,2,\cdots,m\}$,

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}e_i[k]\mathbf{v}[k+1]=0. \tag{6.73}$$

Therefore,

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}\mathbf{e}[k]\mathbf{v}^T[k+1]=0. \tag{6.74}$$

Since $\{\mathbf{z}\}$ also satisfies (6.70), we have,

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}(B\mathbf{e}[k]+\mathbf{w}[k+1]+\mathbf{v}[k+1])(B\mathbf{e}[k]+\mathbf{w}[k+1]+\mathbf{v}[k+1])^T=\sigma_e^2BB^T+\sigma_w^2I_n \tag{6.75}$$

Using (6.74) and the fact that the process noise $\mathbf{w}[k+1]$, and the private excitation of the actuators at time k are independent, the above simplifies to

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}(\mathbf{w}[k+1]\mathbf{v}^T[k+1])+(\mathbf{w}[k+1]\mathbf{v}^T[k+1])^T(\mathbf{v}[k+1]\mathbf{v}^T[k+1])=0. \tag{6.76}$$

Define $S_k \coloneqq \sigma(\mathbf{x}^k, \mathbf{z}^k, \mathbf{e}^{k-2})$, and $\widehat{\mathbf{w}}[k] \coloneqq E[\mathbf{w}[k]\big|S_k]$. Since

$$\mathbf{w}[k] = \mathbf{x}[k] - A\mathbf{x}[k-1] - Bg_{k-1}(\mathbf{z}^{k-1}) - B\mathbf{e}[k-1]),$$

we have

$$(\mathbf{x}^{k-2}, \mathbf{e}^{k-2}) \to (\mathbf{x}[k-1], \mathbf{x}[k], \mathbf{z}^k) \to \mathbf{w}[k]$$

forming a Markov chain. Consequently, $\widehat{\mathbf{w}}[k] \coloneqq E[\mathbf{w}[k]\big|\sigma(\mathbf{x}^{k-2}, \mathbf{e}^{k-2}, \mathbf{x}[k-1], \mathbf{x}[k], \mathbf{z}^k)] = E[\mathbf{w}[k]\big|\sigma(\mathbf{x}[k-1], \mathbf{x}[k], \mathbf{z}^k)]$. Therefore,

$$\widehat{\mathbf{w}}[k] = K_W(B\mathbf{e}[k-1] + \mathbf{w}[k]), \tag{6.77}$$

where $K_W \coloneqq \sigma_w^2(\sigma_e^2 BB^T + \sigma_w^2 I)^{-1}$. Now define $\widetilde{\mathbf{w}}[k] \coloneqq \mathbf{w}[k] - \widehat{\mathbf{w}}[k]$. Then, $(\widetilde{\mathbf{w}}[k-1], \mathcal{S}_k)$ is a martingale difference sequence. This is because $\widetilde{\mathbf{w}}[k-1] \in \mathcal{S}_k$, and

$$E[\widetilde{\mathbf{w}}[k]\big|S_k] = 0. \tag{6.78}$$

Also, $\mathbf{v}[k] \in \mathcal{S}_k$. Hence, MST applies, and we have

$$\sum_{k=0}^{T-1} \widetilde{\mathbf{w}}[k+1]\mathbf{v}^T[k+1] = \begin{bmatrix} o(\sum_{k=0}^{T-1} v_1^2[k+1]) & \dots & o(\sum_{k=0}^{T-1} v_p^2[k+1]) \\ o(\sum_{k=0}^{T-1} v_1^2[k+1]) & \dots & o(\sum_{k=0}^{T-1} v_p^2[k+1]) \\ \vdots & \dots & \vdots \\ o(\sum_{k=0}^{T-1} v_1^2[k+1]) & \dots & o(\sum_{k=0}^{T-1} v_p^2[k+1]) \end{bmatrix} + O(1), \tag{6.79}$$

where $v_i[k+1]$ denotes the i^{th} element of $\mathbf{v}[k+1]$.

Using the above, we have

$$\mathbf{w}[k+1] = \widehat{\mathbf{w}}[k+1] + \widetilde{\mathbf{w}}[k+1] = K_W(B\mathbf{e}[k] + \mathbf{w}[k+1]) + \widetilde{\mathbf{w}}[k+1]. \tag{6.80}$$

From the assumption on the rank of B, it follows that K_W has all eigenvalues strictly lesser than

unity. Therefore, simplifying the above,

$$\mathbf{w}[k+1] = (I - K_W)^{-1} K_W B \mathbf{e}[k] + (I - K_W)^{-1} \widetilde{\mathbf{w}}[k+1]. \tag{6.81}$$

Substituting this into (6.76), using (6.74) and (6.79), and equating the q^{th} entry along the diagonal, we have

$$\sum_{k=0}^{T-1} v_q^2[k+1] + o(\sum_{k=0}^{T-1} v_q^2[k+1]) = o(T). \tag{6.82}$$

Since this is true for all $q \in \{1, 2, ..., p\}$, dividing the above by T and taking the limit as $T \to \infty$ completes the proof. □

6.9 Active Defense for Networked Cyber-Physical Systems: Partially Observed MIMO Systems with Gaussian Process and Measurement Noise

Consider a p^{th} order $m \times n$ partially observed MIMO stochastic linear dynamical system described by

$$\begin{aligned} \mathbf{x}[t+1] &= A\mathbf{x}[t] + B\mathbf{u}[t] + \mathbf{w}[t+1], \\ \mathbf{y}[t+1] &= C\mathbf{x}[t+1] + \mathbf{n}[t+1], \end{aligned} \tag{6.83}$$

where $\mathbf{x}[t] \in \mathbb{R}^p$ is the system's state at time t, $\mathbf{u}[t] \in \mathbb{R}^m$ and $\mathbf{y}[t] \in \mathbb{R}^n$ are respectively the system's input and output at time t, $\mathbf{w}[t] \sim \mathcal{N}(0, Q)$ and $\mathbf{n}[t] \sim \mathcal{N}(0, R)$ with $R > 0$ are respectively the process and observation noises at time t, and A, B, C are known matrices of appropriate dimensions which specify the system dynamics. We assume that the random processes $\{\mathbf{w}\}$ and $\{\mathbf{n}\}$ are independent, and also that each of them is i.i.d. across time.

We denote by $\mathbf{z}[t]$ the measurements reported by the sensors at time t to the controller. A truthful sensor is supposed to report $\mathbf{z} \equiv \mathbf{y}$, but a malicious sensor may report any values for $\{\mathbf{z}\}$.

*Section 6.9 is reprinted with permission from "On minimal tests of sensor veracity for dynamic watermarking-based defense of cyber-physical systems," by B. Satchidanandan and P. R. Kumar in *Proceedings of the 9th International Conference on Communication Systems and Networks (COMSNETS),* pp. 23-30, 2017, IEEE.

We assume the existence of a general history-dependent control policy $\{g_t\}$ according to which the controller computes the input that the actuators should apply at each time t. This control policy is made public, meaning that it is known to all the nodes in the system. While the control policy is supposed to be applied on the actual output sequence $\{\mathbf{y}\}$, since the controller does not directly measure the plant's outputs, it is implemented on $\{\mathbf{z}\}$ as reported by the sensors. Additionally, as outlined in the previous section, in order to secure the system from malicious sensors, the controller commands the actuators to superimpose a private excitation sequence $\{\mathbf{e}\}$ on the sequence of control policy-specified inputs. Hence, the net input applied to the system at time t is given by $u[t] = g_t(\mathbf{z}^t) + \mathbf{e}[t]$, where $\mathbf{e}[t] \sim \mathcal{N}(0, \sigma_e^2 I)$, i.i.d across time, is the controller's private excitation, and $\mathbf{z}^t \coloneqq (\mathbf{z}[0], \mathbf{z}[1], ..., \mathbf{z}[t])$ denotes the past values of $\{\mathbf{z}\}$. Consequently, the system evolves as

$$\begin{aligned} \mathbf{x}[t+1] &= A\mathbf{x}[t] + Bg_t(\mathbf{z}^t) + B\mathbf{e}[t] + \mathbf{w}[t+1], \\ \mathbf{y}[t+1] &= C\mathbf{x}[t+1] + \mathbf{n}[t+1], \end{aligned} \tag{6.84}$$

where $g_t(\mathbf{z}^t) \coloneqq [g_t^1(\mathbf{z}^t)\ , g_t^2(\mathbf{z}^t), ..., g_t^m(\mathbf{z}^t)]^T$.

Assume that (A, C) is observable, and $(A, Q^{\frac{1}{2}})$ is reachable. The controller performs Kalman filtering on the reported sequence of measurements as follows. Let $\mathbf{x}_F(k|k)$ denote the estimate of the state $\mathbf{x}[k]$ given the information upto time k, i.e., $(\mathbf{z}^k, \mathbf{e}^{k-1})$, and $\mathbf{x}_F(k|k-1)$ denote the estimate given the information upto time $k-1$. They are given by the Kalman filtering equations:

$$\mathbf{x}_F(k+1|k+1) = A\mathbf{x}_F(k|k) + Bg_k(\mathbf{z}^k) + B\mathbf{e}[k] + K_{k+1}\boldsymbol{\nu}_F[k+1], \tag{6.85}$$

where $\nu_F[k+1] \coloneqq \mathbf{z}[k+1] - C\mathbf{x}_F(k+1|k)$. We note that if the sensors were truthful, the estimates above would be the conditional mean estimates, and $\boldsymbol{\nu}[t+1]$ would be the innovations process [57] at time t. However, the sensor may be malicious, and so we refer to $\boldsymbol{\nu}_F[t+1]$ as the "false innovations" at time $t+1$. For the purpose of analysis, we also define the "true" Kalman

filter which operates on $\{\mathbf{y}\}$:

$$\mathbf{x}_T(k+1|k+1) = A\mathbf{x}_T(k|k) + Bg_k(\mathbf{z}^k) + B\mathbf{e}[k] + K_{k+1}\boldsymbol{\nu}_T[k+1], \tag{6.86}$$

where $\boldsymbol{\nu}_T[k+1] \coloneqq \mathbf{y}[k+1] - C\mathbf{x}_T(k+1|k)$ is the "true innovations" at time $k+1$. We suppose that the Kalman filters are initialized with the Kalman gain K_0 set to its steady-state value K so that they behave as time-invariant filters [54].

The dynamic watermarking tests we employ are based on the following two observations that hold for the true Kalman filter:

1. $\mathbf{e}[k]$ is independent of $K\boldsymbol{\nu}_T[k+1]$, and

2. the sequence of conditional estimates $\{\mathbf{x}_T(k|k)\}$ satisfies

$$\{\mathbf{x}_T(k+1|k+1) - A\mathbf{x}_T(k|k) - Bg_k(\mathbf{y}^k) - B\mathbf{e}[k]\} \sim \mathcal{N}(0, K\Sigma K^T),$$

where

$$K = PC^T(CPC^T + R)^{-1} \tag{6.87}$$

is the steady-state Kalman gain of the Kalman filter,

$$\Sigma = CPC^T + R \tag{6.88}$$

is the steady-state covariance matrix of the true innovations process, and P is the unique nonnegative definite solution of the discrete algebraic Riccati equation $P = APA^T + Q - APC^T(CPC^T + R)^{-1}CPA^T$. The unique nonnegative definite solution P is guaranteed to exist for the above Riccati equation since (A, C) is observable, $(A, Q^{\frac{1}{2}})$ is reachable, and $R > 0$ [54]. The matrix P has the interpretation of the covariance matrix of the one-step ahead state prediction error of the Kalman filter [54].

The controller therefore performs the following two tests on the reported sequence of observations $\{\mathbf{z}\}$:

1. **Controller Test 1:** Check if the sequence of reported measurements satisfies

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}(\mathbf{x}_F(k+1|k+1)-A\mathbf{x}_F(k|k)-Bg_k(\mathbf{z}^k)-B\mathbf{e}[k])$$
$$(\mathbf{x}_F(k+1|k+1)-A\mathbf{x}_F(k|k)-Bg_k(\mathbf{z}^k)-B\mathbf{e}[k])^T=K\Sigma K^T. \quad (6.89)$$

2. **Controller Test 2:** Check if the sequence of reported observations satisfies

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}\mathbf{e}[k](\mathbf{x}_F(k+1|k+1)-A\mathbf{x}_F(k|k)-Bg_k(\mathbf{z}^k)-B\mathbf{e}[k])^T=0. \quad (6.90)$$

The above tests are equivalent to the tests proposed in [1]. In particular, the second test above can be shown, via straightforward algebraic manipulations, to be equivalent to the corresponding test for first-order SISO systems considered in [1]. We define

$$\mathbf{v}[k+1]\coloneqq\mathbf{x}_F(k+1|k+1)-A\mathbf{x}_F(k|k)-Bg_k(\mathbf{z}^k)-B\mathbf{e}[k]-K\boldsymbol{\nu}_T[k+1], \quad (6.91)$$

and note that if there are no malicious sensors in the system, $\mathbf{v}\equiv 0$.

The following theorem, which is a generalization of the results in [1], establishes the fundamental security guarantee provided by dynamic watermarking.

Theorem 11. *Suppose that (A,C) is observable, $(A,Q^{\frac{1}{2}})$ is reachable, and $R>0$. Further suppose that the matrix CB is of rank n. Then, if the reported measurements $\{\mathbf{z}\}$ pass both (6.90) and (6.89), it can be guaranteed that the additive noise distortion is of zero power, i.e.,*

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}\|\mathbf{v}[k]\|^2=0. \quad (6.92)$$

Proof. We appeal to the following lemma.

Lemma 1: Define $M \coloneqq \Sigma(\Sigma + \sigma_e^2 CBB^T C^T)^{-1}$. If CB is of rank n, then, $(I - M)^{-1}$ exists.

Proof.

$$I - M = I - \Sigma[\Sigma + \sigma_e^2 CBB^T C^T]^{-1}. \tag{6.93}$$

In the above, $[\Sigma + \sigma_e^2 CBB^T C^T]^{-1}$ is guaranteed to exist since $\Sigma > 0$ (because $R > 0$). Its inverse is

$$(I - M)^{-1} = I + \Sigma(\sigma_e^2 CBB^T C^T)^{-1}, \tag{6.94}$$

since $(\sigma_e^2 CBB^T C^T)^{-1}$ exists because CB has rank n. □

Since the reported measurements pass (6.90), we have using (6.91),

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} \mathbf{e}[k](K\boldsymbol{\nu}_T[k+1] + \mathbf{v}[k+1])^T = 0.$$

Since $\mathbf{e}[k]$ is independent of the innovations $\boldsymbol{\nu}_T[k+1]$, the above simplifies to

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} \mathbf{e}[k]\mathbf{v}^T[k+1] = 0. \tag{6.95}$$

Since the reported measurements also pass (6.89), we have using (6.91),

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (K\boldsymbol{\nu}_T[k+1] + \mathbf{v}[k+1])(K\boldsymbol{\nu}_T[k+1] + \mathbf{v}[k+1])^T = K\Sigma K^T.$$

Simplifying the above gives

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} K\boldsymbol{\nu}_T[k+1]\mathbf{v}^T[k+1] + (K\boldsymbol{\nu}_T[k+1]\mathbf{v}^T[k+1])^T + \mathbf{v}[k+1]\mathbf{v}^T[k+1] = 0. \tag{6.96}$$

Define the σ-algebra $\mathcal{S}_{k+1} \coloneqq \sigma(\mathbf{y}^{k+1}, \mathbf{z}^{k+1}, \mathbf{e}^{k-1}, \mathbf{x}_T^{k|k})$, where $\mathbf{x}_T^{k|k} \coloneqq (\mathbf{x}_T(0|0), \mathbf{x}_T(1|1), ..., \mathbf{x}_T(k|k))$.

We also define $\widehat{\boldsymbol{\nu}}_T[k] \coloneqq \mathbb{E}[\boldsymbol{\nu}_T[k]\big|\mathcal{S}_k]$, and $\widetilde{\boldsymbol{\nu}}_T[k] \coloneqq \boldsymbol{\nu}_T[k] - \widehat{\boldsymbol{\nu}}_T[k]$. Then, from the definition of the innovations at time $k+1$, we have

$$\boldsymbol{\nu}_T[k+1] \coloneqq \mathbf{y}[k+1] - C\mathbf{x}_T(k+1|k) = \mathbf{y}[k+1] - CA\mathbf{x}_T(k|k) - CBg_k(\mathbf{z}^k) - CB\mathbf{e}[k]. \tag{6.97}$$

From the above, it follows that

$$(\mathbf{y}^k, \mathbf{e}^{k-1}, \mathbf{x}_T^{k-1|k-1}) \to (\mathbf{x}_T(k|k), \mathbf{y}[k+1], \mathbf{z}^{k+1}) \to \boldsymbol{\nu}_T[k+1]$$

forms a Markov chain. Therefore,

$$\widehat{\boldsymbol{\nu}}_T[k+1] \coloneqq \mathbb{E}[\boldsymbol{\nu}_T[k+1]\big|\sigma(\mathbf{y}^{k+1}, \mathbf{z}^{k+1}, \mathbf{e}^{k-1}, \mathbf{x}_T^{k|k})] = \mathbb{E}[\boldsymbol{\nu}_T[k+1]\big|\sigma(\mathbf{x}_T(k|k), \mathbf{y}[k+1], \mathbf{z}^{k+1})].$$

Combining the above with (6.97), we have the MMSE estimate as [54, Chapter 7, Lemma 2.5]

$$\widehat{\boldsymbol{\nu}}_T[k+1] = M(CB\mathbf{e}[k] + \boldsymbol{\nu}_T[k+1]). \tag{6.98}$$

Hence,

$$\boldsymbol{\nu}_T[k+1] = \widehat{\boldsymbol{\nu}}_T[k+1] + \widetilde{\boldsymbol{\nu}}_T[k+1] = MCB\mathbf{e}[k] + M\boldsymbol{\nu}_T[k+1] + \widetilde{\boldsymbol{\nu}}_T[k+1]$$

Rearranging and using Lemma 1, we have

$$\boldsymbol{\nu}_T[k+1] = (I-M)^{-1}MCB\mathbf{e}[k] + (I-M)^{-1}\widetilde{\boldsymbol{\nu}}_T[k+1]. \tag{6.99}$$

Now, the RHS of (6.97), and hence $\boldsymbol{\nu}_T[k+1]$, is measurable with respect to $\mathcal{S}_{k+2}$. Also, since $\widehat{\boldsymbol{\nu}}_T[k+1] \in \mathcal{S}_{k+1} \subset \mathcal{S}_{k+2}$, it follows that $\widetilde{\boldsymbol{\nu}}_T[k+1] \in \mathcal{S}_{k+2}$. Clearly, $\mathbb{E}[\widetilde{\boldsymbol{\nu}}_T[k+2]\big|\mathcal{S}_{k+2}] = 0$. Hence, we have that $(\widetilde{\boldsymbol{\nu}}_T[k+1], \mathcal{S}_{k+2})$ is a martingale difference sequence. Moreover, since $\mathbf{v}[k+1]$, after

some algebra, can be expressed as $K(\mathbf{z}[k+1] - \mathbf{y}[k+1]) - KCA(\mathbf{x}_F(k|k) - \mathbf{x}_T(k|k))$, we have $\mathbf{v}[k+1] \in \mathcal{S}_{k+1}$. Hence, Martingale Stability Theorem (MST) [49, Lemma 2(iii)] holds, and we have

$$\sum_{k=0}^{T-1} \widetilde{\boldsymbol{\nu}}_T[k+1]\mathbf{v}^T[k+1] = \begin{bmatrix} o(\sum_{k=1}^T v_1^2[k]) & \dots & o(\sum_{k=1}^T v_p^2[k]) \\ \vdots & \vdots & \vdots \\ o(\sum_{k=1}^T v_1^2[k]) & \dots & o(\sum_{k=1}^T v_p^2[k]) \end{bmatrix} + [O(1)]_{p\times p}, \tag{6.100}$$

where $[O(1)]_{p\times p}$ denotes a $p \times p$ matrix all of whose entries are $O(1)$. Substituting (6.99) in (6.96) and using (6.95) and (6.100) yields

$$\sum_{k=1}^{T} \mathbf{v}[k]\mathbf{v}^T[k] + \begin{bmatrix} o(T) & \dots & o(T) \\ \vdots & \vdots & \vdots \\ o(T) & \dots & o(T) \end{bmatrix} + \begin{bmatrix} o(\sum_{k=1}^T v_1^2[k]) & \dots & o(\sum_{k=1}^T v_p^2[k]) \\ \vdots & \vdots & \vdots \\ o(\sum_{k=1}^T v_1^2[k]) & \dots & o(\sum_{k=1}^T v_p^2[k]) \end{bmatrix}$$
$$+ \begin{bmatrix} o(\sum_{k=1}^T v_1^2[k]) & \dots & o(\sum_{k=1}^T v_1^2[k]) \\ \vdots & \vdots & \vdots \\ o(\sum_{k=1}^T v_p^2[k]) & \dots & o(\sum_{k=1}^T v_p^2[k]) \end{bmatrix} = o(T). \tag{6.101}$$

Dividing the above by T, equating the trace, and letting $T \to \infty$ completes the proof. □

We now show that if one drops either of the two controller tests (6.90) or (6.89), then the guarantee does not hold. We do so by explicitly constructing two attack strategies, each of which passes exactly each one of the tests, and yet, $\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^T \|\mathbf{v}[k]\|^2 \neq 0$ for both the attacks.

Consider a special case of system (6.83), viz., a SISO first-order perfectly observed system ($p = m = n = C = 1, R = 0, Q = \sigma_w^2 \in \mathbb{R}_+$). In that case, $\mathbf{x}_F(k|k)$ reduces to $\mathbf{z}[k]$, $\boldsymbol{\nu}_T[k]$ to $w[k]$, K to 1, and Σ to σ_w^2. Consequently, tests (6.89) and (6.90) reduce respectively to

1. **Test 1:** Check if the reported sequence of measurements $\{z\}$ satisfies

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}(z[k+1]-Az[k]-Bg_k(z^k)-Be[k])^2=\sigma_w^2, \tag{6.102}$$

2. **Test 2:** Check if the reported sequence of measurements $\{z\}$ satisfies

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}e[k](z[k+1]-Az[k]-Bg_k(z^k)-Be[k])=0, \tag{6.103}$$

and $v[t+1]=z[t+1]-Az[t]-Bg_t(z^t)-Be[t]-w[t+1]$.

Counterexample showing Controller Test 1 alone is not sufficient: Suppose that the reported measurements are subjected to (6.102) alone, the first test. To show that this is not sufficient to guarantee zero additive noise distortion power by the malicious sensor, consider the following attack.

Suppose that the malicious sensor reports measurements $\{z\}$ generated as

$$z[k+1]=Az[k]+Bg_k(z^k)+(\frac{B^2\sigma_e^2-\sigma_w^2}{B^2\sigma_e^2+\sigma_w^2})(y[k+1]-Ay[k]-Bg_k(z^k)). \tag{6.104}$$

We now show that the sequence $\{z\}$ so generated passes (6.102), the first test.

Define $\gamma[k]\coloneqq z[k]-Az[k-1]-Bu^g[k-1]-Be[k-1]$, the quantity whose second moment is being empirically tested in (6.102). Then, we have

$$\begin{aligned}\gamma[k]&=z[k]-Az[k-1]-Bg_{k-1}(z^{k-1})-Be[k-1]\\&=(\frac{B^2\sigma_e^2-\sigma_w^2}{B^2\sigma_e^2+\sigma_w^2})(y[k]-Ay[k-1]-Bg_{k-1}(z^{k-1})-Be[k-1])\\&=(\frac{B^2\sigma_e^2-\sigma_w^2}{B^2\sigma_e^2+\sigma_w^2})(Be[k-1]+w[k])-Be[k-1]\\&=-\frac{2\sigma_w^2}{B^2\sigma_e^2+\sigma_w^2}Be[k-1]+(\frac{B^2\sigma_e^2-\sigma_w^2}{B^2\sigma_e^2+\sigma_w^2})w[k].\end{aligned} \tag{6.105}$$

From the above, it is clear that $\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}\gamma^2[k]$ of (6.102) is simply the variance of the RHS

of the above, given by

$$(\frac{-2\sigma_w^2 B}{B^2\sigma_e^2+\sigma_w^2})^2 + (\frac{B^2\sigma_e^2-\sigma_w^2}{B^2\sigma_e^2+\sigma_w^2})^2.$$

This simplifies to σ_w^2, and hence, this attack passes Test 1.

Finally, for the above attack, it is easy to see that $v[k+1] = \gamma[k+1] - w[k+1] = -\frac{2\sigma_w^2}{B^2\sigma_e^2+\sigma_w^2}(Be[k]+w[k+1])$, and hence, $\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T} v^2[k] = \frac{4\sigma_w^4}{B^2\sigma_e^2+\sigma_w^2} \neq 0$. $\square$

Counterexample showing Controller Test 2 alone is not sufficient: Now suppose that the reported measurements are subjected to (6.103) alone. To show that this is not sufficient to guarantee zero additive noise distortion power by the malicious sensor, consider the following attack. The sensor reports measurements $\{z\}$ generated as

$$z[k+1] = Az[k] + Bg_k(z^k) + (y[k+1] - Ay[k] - Bg_k(z^k) + \lambda[k+1]), \tag{6.106}$$

where $\lambda[k+1] \sim \mathcal{N}(0,\sigma_\lambda^2)$ is chosen by the sensor in an i.i.d. fashion across time, and also independently of all random variables that it has observed till then.

To show that the sequence $\{z\}$ so generated passes (6.103), the second test, note that

$$\begin{aligned}
&\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1} e[k](z[k+1] - Az[k] - Bg_k(z^k) - Be[k]) \\
=&\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1} e[k](y[k+1] - Ay[k] - Bg_k(z^k) + \lambda[k+1] - Be[k]) \\
=&\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1} e[k](w[k+1] + \lambda[k+1]).
\end{aligned}$$

Since $w[k+1]$ and $\lambda[k+1]$ are independent of $e[k]$, the above reduces to 0, thereby passing (6.103), and hence, (6.90).

However, for the above attack, it is easy to see that $v[k+1] = \lambda[k+1]$, and hence,

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T} v^2[k] = E[\lambda^2[k]] = \sigma_\lambda^2 \neq 0. \quad \square$$

Remark: It is well known that the innovations process of a stochastic process is a causal and causally invertible transformation of the stochastic process with the property that it is uncorrelated across time [57]. Hence, the innovations at time t can be thought of as summarizing all the "new" information provided by the sensors at time t, information that could not have been predicted from the past. Therefore, one can think of the honest sensors' purpose as being to report the innovations at each time t. Now, from (6.91), we have $\mathbf{x}_F(k+1|k+1) - A\mathbf{x}_F(k|k) - Bg_k(\mathbf{z}^k) - B\mathbf{e}[k] = K\boldsymbol{\nu}_T[k+1] + \mathbf{v}[k+1]$. The LHS of the above can be computed by the controller. Hence, $\mathbf{v}[k+1]$ has a physical interpretation as the distortion added by the malicious sensors to the true innovations at time $k+1$ (hence the nomenclature for additive noise distortion power). What the above theorem says is that the malicious sensors cannot distort the true innovations process beyond adding a zero-power sequence to it if they wish to remain undetected.

We now consider linear control designs that provide some guarantee on the quadratic state tracking error. The design need not be an optimal LQG design, but one that merely aims at providing some upper bound on the aforementioned quantity. The following theorem shows that for stable systems, guaranteeing that any additive noise distortion is of power zero is sufficient to ensure that the malicious sensors do not increase the quadratic cost of the state from its design value in the case of linear designs.

Theorem 12. *Suppose that the system (6.83) is open-loop stable, i.e., A has all its eigenvalues in the open left half-plane, and define*

$$\mathbf{d}[k] \coloneqq \mathbf{x}_F(k|k) - \mathbf{x}_T(k|k). \tag{6.107}$$

If the reported measurements pass (6.90) and (6.89), then,

1.

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} \|\mathbf{d}[k]\|^2 = 0. \tag{6.108}$$

2. *Suppose that the control policy is a linear feedback policy $g_t(\mathbf{z}^t) = F\mathbf{x}_F(t|t)$, and a control objective is quadratic regulation. Then, the true quadratic regulation performance $\lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} \|\mathbf{x}_T(k|k)\|^2$ of the system is no different from what the controller thinks it is, in the sense that*

$$\lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} \|\mathbf{x}_F(k|k)\|^2 = \lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} \|\mathbf{x}_T(k|k)\|^2 \tag{6.109}$$

3. *Under the same conditions as (2) above, the malicious sensors cannot increase the quadratic regulation cost of the system $\lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} \|\mathbf{x}_T(k|k)\|^2$ from the value that would be obtained if all the sensors were honest.*

Proof. Subtracting (6.86) from (6.85), we have $\mathbf{d}[k+1] = A\mathbf{d}[k] + K(\boldsymbol{\nu}_F[k+1] - \boldsymbol{\nu}_T[k+1])$. From (6.85), we have $K\boldsymbol{\nu}_F[k+1] = \mathbf{x}_F(k+1|k+1) - A\mathbf{x}_F(k|k) - Bg_k(\mathbf{z}^k) - B\mathbf{e}[k]$. Combining this with (6.91) gives $K(\boldsymbol{\nu}_F[k+1] - \boldsymbol{\nu}_T[k+1]) = \mathbf{v}[k+1]$. Substituting this in the above equation gives $\mathbf{d}[k+1] = A\mathbf{d}[k] + \mathbf{v}[k+1]$. Since $\{\mathbf{v}\}$ is of zero power and A is stable, result (1) follows.

Now, from (6.107), we have $\mathbf{x}_F(k|k) = \mathbf{x}_T(k|k) + \mathbf{d}[k]$. By triangular inequality, we have $\|\mathbf{x}_F(k|k)\| \le \|\mathbf{x}_T(k|k)\| + \|\mathbf{d}[k]\|$. Hence, $\|\mathbf{x}_F(k|k)\|^2 \le \|\mathbf{x}_T(k|k)\|^2 + \|\mathbf{d}[k]\|^2 + 2\|\gamma\mathbf{x}_T(k|k)\|\|\gamma^{-1}\mathbf{d}[k]\|$ for all $\gamma > 0$. Since $2\|\gamma\mathbf{x}_T(k|k)\|\|\gamma^{-1}\mathbf{d}[k]\| \le \|\gamma\mathbf{x}_T(k|k)\|^2 + \|\gamma^{-1}\mathbf{d}[k]\|^2$, substituting this in the above yields $\|\mathbf{x}_F(k|k)\|^2 \le (1+\gamma^2)\|\mathbf{x}_T(k|k)\|^2 + (1+\gamma^{-2})\|\mathbf{d}[k]\|^2$. Hence,

$$\lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} \|\mathbf{x}_F(k|k)\|^2 \le \lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} (1+\gamma^2)\|\mathbf{x}_T(k|k)\|^2 + \lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} (1+\gamma^{-2})\|\mathbf{d}[k]\|^2$$

The second term reduces to zero from the previous result. Since the above is true for all $\gamma > 0$, taking $\gamma \to 0$ gives

$$\lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} \|\mathbf{x}_F(k|k)\|^2 \le \lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} \|\mathbf{x}_T(k|k)\|^2. \tag{6.110}$$

Similarly, from (6.107), we have $\mathbf{x}_T(k|k) = \mathbf{x}_F(k|k) - \mathbf{d}[k]$. Hence, $\|\mathbf{x}_T(k|k)\| = \|\mathbf{x}_F(k|k) -$

$\mathbf{d}[k]\| \leq \|\mathbf{x}_F(k|k)\| + \|\mathbf{d}[k]\|$. Continuing as above, we arrive at

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} \|\mathbf{x}_T(k|k)\|^2 \leq \lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} \|\mathbf{x}_F(k|k)\|^2. \tag{6.111}$$

Combining the above with (6.110) gives the second result.

It follows from the above result that even though the controller does not have access to the true measurements $\{\mathbf{y}\}$, it can empirically compute the true quadratic regulation cost $\mathbb{E}[\|\mathbf{x}_T(k|k)\|^2]$ of the system. It follows that the malicious sensors cannot increase the true quadratic regulation cost of the system from its design value without exposing their presence. □

6.10 Extension to Non-Gaussian Systems

The results developed in the previous sections assumed that the process noise follows a Gaussian distribution. This assumption can be relaxed to a certain extent. In this section, we illustrate how this may be relaxed for a single-input-single-output system.

Consider a SISO system described by

$$x[t+1] = ax[t] + bu[t] + w[t], \tag{6.112}$$

where $w[t] \sim P_W$ is an i.i.d. process with mean 0 and variance σ_w^2. In such a case, the actuator can choose the distribution of its private excitation such that the output of the private excitation has the same distribution as the process noise. Hence, the actuator applies to the system the input

$$u[t] = g_t(z^t) + e[t], \tag{6.113}$$

where $be[t] \sim P_W$ is an i.i.d. sequence. As before, even though the actuator implements the control policy $\{g\}$, the policy is applied to the measurements $z[t]$ reported by the sensor, which could differ

from the true output $x[t]$. Therefore, the system evolves in closed-loop as

$$x[t+1] = ax[t] + bg_t(z^t) + be[t] + w[t]. \tag{6.114}$$

The actuator then performs the following tests to check if the sensor is malicious or not.

1) **Actuator Test 1:** Check if the reported sequence of measurements $\{z[t]\}$ satisfies

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (z[k+1] - az[k] - bg_k(z^k) - be[k])^2 = \sigma_w^2. \tag{6.115}$$

2) **Actuator Test 2:** Check if the reported sequence of measurements $\{z[t]\}$ satisfies

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (z[k+1] - az[k] - bg_k(z^k))^2 = 2\sigma_w^2. \tag{6.116}$$

As before, let

$$v[t+1] \coloneqq z[t+1] - az[t] - bg_t(z^t) - be[t] - w[t+1],$$

so that for an honest sensor which reports $z[t] \equiv x[t]$, $v[t] = 0 \ \ \forall t$. The *additive noise distortion power* of a malicious sensor too is defined as before, i.e., as

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} v^2[k].$$

The following result, a generalization of Theorem 1, shows that the above tests suffice to ensure that a malicious sensor of only zero effective power can remain undetected.

Theorem 13. *If $\{z[t]\}$ satisfies tests (6.115) and (6.116), then,*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} v^2[k] = 0. \tag{6.117}$$

Proof. Following the same sequence of arguments as in the proof of Theorem 1, we arrive at the

following equalities:

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}v^2[k]+\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}2v[k]w[k]=0, \tag{6.118}$$

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}e[k-1]v[k]=0. \tag{6.119}$$

Let $\mathcal{S}_k \coloneqq \sigma(x^k, z^k, e^{k-2})$, and $\widehat{w}[k] \coloneqq E[w[k]\big|\mathcal{S}_k]$. The general complication is that due to non-Gaussianity, the conditional mean estimate may be non-linear in $e[k-1]$ and $w[k]$, unlike in the Gaussian case. However, since the sequence of observations are i.i.d., and the distribution of the actuator's private excitation is chosen to be the same as the distribution of the process noise, we have

$$\widehat{w}[k]=\frac{1}{2}(be[k-1]+w[k])=\beta(be[k-1]+w[k]), \tag{6.120}$$

where $\beta \coloneqq \frac{1}{2}$. This can be written as

$$\widehat{w}[k]=\alpha e[k-1]+\beta w[k],$$

where $\alpha \coloneqq b\beta$. Let $\widetilde{w}[k] \coloneqq w[k]-\widehat{w}[k]$.

Now, the following result also holds following the same sequence of arguments employed in the proof of Theorem 1:

$$\sum_{k=1}^{T}v[k]\widetilde{w}[k]=o(\sum_{k=1}^{T}v^2[k])+O(1). \tag{6.121}$$

Hence, following similar arguments as in the proof of Theorem 1, we obtain

$$\sum_{k=1}^{T}v^2[k]+\sum_{k=1}^{T}2v[k]w[k]=(1+o(1))(\sum_{k=1}^{T}v^2[k])+O(1).$$

Dividing the above equation by T, taking the limit as $T \to \infty$, and invoking (6.118) completes the proof. □

Remark: In the watermarking strategy used in the above system, in order to safeguard against malicious actions of the sensors, the level of noise added is exactly equal to the process noise already present in the system. Hence there is an amplification of the process noise's standard deviation by $\sqrt{2}$ that appears the price of guarding against malicious attack by the sensor. In contrast, in the Gaussian case, we have seen that this extra cost can be made as small as desired by choosing σ_e^2 as small as desired, though of course at the cost of delaying detection with an acceptable false alarm probability, as we discuss in the next section. In Chapter 7, we will characterize the set of all watermark strategies that can guarantee (6.117).

6.11 Statistical Tests for Active Defense

The results developed in the previous sections are couched in an asymptotic fashion. They characterize what can be detected and how that may be done. These asymptotic characterizations can be used to develop statistical tests that reveal malicious activity in a finite period of time with acceptable false alarm rates.

The task at hand is to develop statistical tests equivalent to asymptotic tests such as (6.12), (6.13), (6.70), (6.71), which include tests for covariance matrices of certain random vectors. This problem has been addressed in [58] in the context of fault detection in control systems, in which the innovation sequence is tested for the properties of whiteness, mean, and covariance. Since these are also the properties that have to be satisfied by the test sequences that we construct in (6.12), (6.13), and (6.71), the same test described in [58] can be used.

Yet another test that is particularly well-suited for the problem at hand is the sequential probability ratio test, introduced in [59]. In the standard setting of sequential hypothesis testing, for each time t, the set of all possible observations S_m (or the space of sufficient statistics) is partitioned into three sets R_0, R_1, R. At each time t, a decision is made whether to select the null hypothesis (and stop), select the the alternative hypothesis (and stop), or continue observing. The null (alternative)

hypothesis is accepted and hypothesis testing is stopped if at some time t, the observations available till time t, denoted by $\{z_s^t\}$, fall in the set R_0 (R_1). Hypothesis testing is continued at time t if $\{z_s^t\}$ falls in R. One such sequential test is the sequential probability ratio test presented in [59], where the likelihood ratio of observations up to time t, denoted by Λ_t, is compared against two thresholds τ_1 and τ_2. The null (alternative) hypothesis is accepted at time t if $\Lambda_t < \tau_1$ ($\Lambda_t > \tau_2$). If $\tau_1 \leq \Lambda_t \leq \tau_2$, another observation is drawn, and the process repeats.

There are a few aspects where the problem at hand departs from the above. These are enumerated below. In what follows, the null hypothesis is that the control system is not under attack.

1) In the problem of secure control, it is not possible to attribute a distribution (or even a parameterized, finite-dimensional family of distributions) from which the observations would occur under the alternative hypothesis (that the control system is under attack). Consequently, a likelihood ratio cannot be defined, and the sequential probability ratio test cannot be used. Hence, the problem at hand is not to test one hypothesis against another, but is only that of rejecting the null hypothesis or otherwise.

2) In the classical setup as described in [59], in a finite time and with probability 1, the test stops, and one hypothesis or the other is accepted. However, the adversary could be a "sleeper agent". It could behave as an honest party for a long period of time, and then operate in a malicious manner. Hence in our situation, at no time can one "accept" the null hypothesis forever for all subsequent time. Therefore, in the systems of interest here, there are only two possible decisions at any time t, reject the null hypothesis or continue observations. An alternative, more positive way of stating this, is that the watermarking is ever vigilant to detecting attacks, and never precludes the possibility of a future attack.

3) In the classical setting, optimal performance is obtained by considering all observations that are available up to the time of making a decision. However, in our problem, an adversary that behaves maliciously in a bursty fashion, is not likely to be detected since long term averaging would nullify these effects. In order to account for this, a moving window approach or

exponential forgetting [54] could be employed over which a test statistic can be calculated. In the moving window approach, the statistical test is conducted over a window of l observations, $nl \leq t < (n+1)l$. In each window it assesses whether to reject the null hypotheses. In this manner, malicious activity can be detected within l samples, with acceptable false alarm rate. The choice of the excitation variance σ_e^2 can be based on the acceptable detection delay l, and the false alarm rate. In the exponential forgetting approach, the test discounts past values by a geometric sequence. Its advantage is that it allows a recursive implementation that does not require the storing of a window of observations which may be a matter of concern in electronic control units or processors of limited memory. Another advantage is that just like the window length, the forgetting factor λ can be varied, with its value of λ corresponding roughly to a window length of $9/(1-\lambda)$.

Based on the above, the basis of a statistical hypothesis test for the problem can be chosen as follows. Under the null hypothesis, the sample covariance matrix follows the Wishart distribution [60] (the multidimensional counterpart of the Chi-squared distribution). Hence, for a given false alarm rate α, the threshold $\tau(\alpha)$ can be determined for the likelihood function of the observations, where the likelihood function considered for the test at time t is the probability density evaluated at the l most recent observations. If this likelihood function exceeds the threshold at any time t, an alarm is raised. If not, the test is repeated for the next time instant. The following simulation example is illustrative.

Example: We consider the scenario of an ARX system addressed in Section-6.5. Specifically, we consider a second order plant of the type common in process control. A single-input, single-output stable plant obeying (6.25), with parameters $a_0 = 0.7$, $a_1 = 0.2$, $b_0 = 1$, $b_1 = 0.5$ and $\sigma_w^2 = 1$ is considered. Note that the chosen system is minimum phase, so that the pre-equalizer required to filter the sequence of private excitation is stable.

The sensor is assumed to be initially well-behaved, but suddenly become malicious after a certain period (denoting the time of initiation of attack), while the actuator is assumed to remain

uncompromised. The actuator applies control inputs

$$u[k] = -\frac{1}{b_0}(a_0 z[t] + a_1 z[t-1] + b_1 u[k-1]) + e[k],$$

where $z[k]$ is the measurement reported by the sensor at time k, and $e[k] \sim$ i.i.d. $\mathcal{N}(0,1)$ is the actuator node's private excitation. Therefore, the system evolves as

$$y[k+1] = 0.7(y[k] - z[k]) + 0.3(y[k-1] - z[k-1]) + e[k] + w[k+1].$$

The adversary attacks the system at some random time and begins to report false measurements. In order to do so, the adversary estimates optimally the process noise at each instant from the measurements that it observes. Since the process noise and the private excitation have the same variance, the optimal estimate is simply $\frac{1}{2}(y[k+1] - 0.7(y[k] - z[k]) - 0.3(y[k-1] - z[k-1])$. Once the adversary forms its estimate, it adds a noise correlated with the estimate as follows. The adversary forms $v[t] = n[t] - \widehat{w}[t]$, where $n[t] \sim \mathcal{N}(0,1)$, the same distribution as $w[t]$, and reports $z[t] = x[t] + v[t]$. Note that in the absence of dynamic watermarking, employing this strategy would enable the adversary to form perfect estimates of the process noise, and consequently, as reasoned in Section-6.3, would render every detection algorithm useless.

Fig. 6.3 plots the evolution of the negative log likelihood function over time when dynamic watermarking is employed. In our simulation, the adversary initiates attack at time epoch 4500. As indicated in Fig. 6.3, the (windowed) negative log likelihood function corresponding to (6.29) stays within limits until the attack. Around time epoch 4500, the likelihood function steadily increases, indicating the onset of an attack. Based on tolerable false alarm rates, an appropriate detection threshold can be set.

6.12 Concluding Remarks

In this chapter, we have considered the problem of securing a cyberphysical system from malicious sensors and have developed a general procedure termed "dynamic watermarking" that im-

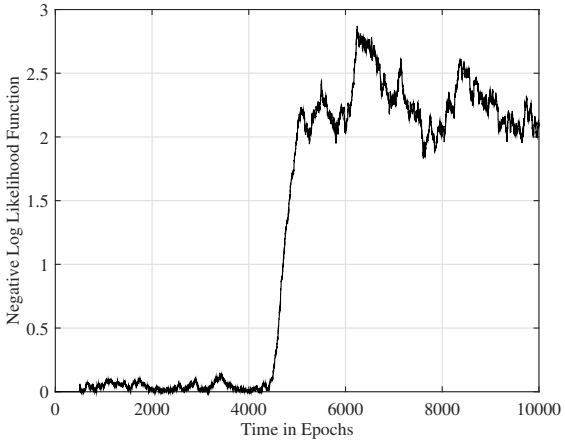

Figure 6.3: Sequential Hypothesis Testing for Attack Detection. Reprinted with permission from [1].

poses private excitation signals on the actuation signals whose presence can be traced around the loop to detect malicious behaviors of sensors. We have illustrated the technique in the context of a variety of control system contexts including a simple first-order SISO system, ARX and ARMAX systems, a staple model in industrial process control, and the partially-observed MIMO system with process and measurement noise, one of the most general system models used in modern control theory. We have established the fundamental security guarantee provided by dynamic watermarking in all of the aforementioned system contexts. These guarantees are contingent on the controller conducting two particular tests of sensor veracity, and it was shown via explicit construction of two attack strategies that both of these tests are required in that neither can be dropped if the security guarantees are to hold.

Addendum

Notation: This chapter uses the following notation.

- Scalars are denoted using lowercase: a
- Vectors are denoted using lowercase boldface: $\mathbf{x}$
- The i^{th} component of vector $\mathbf{x}$: x_i

- Matrices are denoted using uppercase: A
- The i^{th} column of matrix A: $A_{\cdot i}$
- The i^{th} row of matrix A: $A_{i\cdot}$
- The submatrix of A formed by the intersection of rows i through j and columns p through q: $A_{i:j,p:q}$
- The matrix with i^{th} column of A removed: $A_{\cdot(-i)}$
- The matrix with i^{th} row of A removed: $A_{(-i)\cdot}$
- The vector with i^{th} component of $\mathbf{x}$ removed: $\mathbf{x}_{-i}$

7. ON THE DESIGN OF SECURITY-GUARANTEEING DYNAMIC WATERMARKS*

7.1 Introduction

In this chapter, we go beyond the class of Gaussian systems and address the problem of designing watermarks for systems with arbitrarily distributed additive noise. Specifically, given a perfectly observed, Multiple-Input-Multiple-Output (MIMO), linear stochastic system with independent, identically distributed (i.i.d.) arbitrary process noise distribution, we derive the necessary and sufficient conditions that the watermark should satisfy in order for the fundamental security guarantee to hold.

The results of this chapter characterize the feasible set of watermarks that provide the fundamental security guarantee for any given noise distribution. Obtaining this set is a prerequisite for designing optimal watermarking schemes that optimize desirable objectives such as minimizing the watermark power, minimizing the maximum value that the watermark can assume, etc.

The rest of this chapter is organized as follows. Section 7.2 formulates the problem. Section 7.3 contains the main result of the chapter, viz., the necessary and sufficient conditions that the statistics of the watermark should satisfy in order for the fundamental security guarantee to hold. Section 7.4 contains some concluding remarks.

Notation and terminology: Given a vector $\mathbf{x}$, we denote by x_i the i^{th} entry of $\mathbf{x}$, and by $\mathbf{x}^{TT}$ the matrix $\mathbf{x}\mathbf{x}^T$. The letter Ψ always denotes a conditional expectation. Given two random variables X and Y, we denote by $\Psi^X(Y)$ the conditional expectation $\mathbb{E}[X|Y]$. When clear from the context, we drop the argument Y and denote the above conditional expectation by simply Ψ^X. The set of random variables with finite second moments forms a Hilbert space with $\mathbb{E}[XY]$ being defined as the inner product between random variables X and Y. The projection operations on random variables that are described in this chapter are all with respect to this inner product.

*Reprinted with permission from "On the Design of Security-Guaranteeing Dynamic Watermarks," by B. Satchidanandan and P. R. Kumar, *IEEE Control Systems Letters*, 2019, IEEE.

7.2 Motivation and Problem Formulation

Consider a p^{th} order, perfectly observed, multiple-input-multiple-output linear stochastic system described by

$$\mathbf{x}[t+1] = A\mathbf{x}[t] + B\mathbf{u}[t] + \mathbf{w}[t+1], \tag{7.1}$$

where $\mathbf{x}[t] \in \mathbb{R}^p$ is the state as well as the output of the system at time t, $\mathbf{u}[t] \in \mathbb{R}^m$ is the input applied to the system at time t, $\mathbf{w}[t+1] \in \mathbb{R}^p$ is the process noise affecting the state at time $t+1$, and A and B are known matrices of appropriate dimensions. The process noise sequence $\{\mathbf{w}\}$ is i.i.d. across time, and we denote by P_W the probability distribution of $\mathbf{w}[t]$ at each time t. We denote by $\boldsymbol{\mu}_W$ the mean of the process noise, and by Σ_W the matrix $\mathbb{E}(\mathbf{w}[t]\mathbf{w}^T[t])$. Apart from the assumption that these two moments exist, no other assumption is made on the distribution P_W.

The input at each time t is determined by a controller according to an arbitrary, possibly history-dependent control policy $\{g_t\}$, where $g_t : \mathbb{R}^{p\times(t+1)} \to \mathbb{R}^m$ is the function that maps the measurement sequence reported by the sensors up to time t to the input applied by the actuators at time t.

The problem setup that we consider in this chapter is one where an arbitrary subset of the sensors in system (7.1) could be malicious. The malicious sensors may not report the measurements that they observe in a truthful manner. Rather, they may distort the measurements that they observe in an arbitrary fashion, possibly according to some coordinated attack strategy, and report the distorted measurements to the controller.

In order to defend against malicious sensors that may be present in the system, the controller at each time t superimposes on the control policy-specified input a random signal $\mathbf{e}[t]$, known as the watermark, chosen independently of all other random variables in the system up until that time, and in an i.i.d. fashion across time. We denote by P_E the probability distribution of $\mathbf{e}[t]$ at each time t. We constrain the watermark to have finite second moment so that its mean and covariance matrix exist. We denote by $\boldsymbol{\mu}_E$ its mean and by Σ_E the matrix $\mathbb{E}(\mathbf{e}[t]\mathbf{e}^T[t])$. The net input applied

to the system at time t is $\mathbf{u}[t] = g_t(\mathbf{z}^t) + \mathbf{e}[t]$, and the system (7.1) evolves in closed loop as

$$\mathbf{x}[t+1] = A\mathbf{x}[t] + Bg_t(\mathbf{z}^t) + B\mathbf{e}[t] + \mathbf{w}[t+1], \tag{7.2}$$

where $\mathbf{z}^t \coloneqq \{\mathbf{z}[0], \ldots, \mathbf{z}[t]\}$.

Define $\Sigma_{EW} \coloneqq \mathbb{E}(\mathbf{e}[t]\mathbf{w}^T[t+1])$. In order to detect whether or not there are any malicious sensors in the system, the controller conducts the following tests, and declares the presence of malicious sensors if any of them fail.

Test 1: The controller checks if

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (\mathbf{z}[k+1] - A\mathbf{z}[k] - Bg_k(\mathbf{z}^k))^{TT} \stackrel{?}{=} B\Sigma_E B^T + \Sigma_W + B\Sigma_{EW} + \Sigma_{EW}^T B^T. \tag{7.3}$$

Test 2: The controller checks if

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} \mathbf{e}[k]\ (\mathbf{z}[k+1] - A\mathbf{z}[k] - Bg_k(\mathbf{z}^k) - B\mathbf{e}[k])^T \stackrel{?}{=} \Sigma_{EW}. \tag{7.4}$$

Test 3: The controller checks if

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (\mathbf{z}[k+1] - A\mathbf{z}[k] - Bg_k(\mathbf{z}^k) - B\mathbf{e}[k]) \stackrel{?}{=} \boldsymbol{\mu}_W. \tag{7.5}$$

While the tests are presented in an asymptotic form, as illustrated in [61], they can be converted into finite-time statistical tests with desired false alarm and misdetection rates via standard methods using thresholds on the finite time running averages. In comparison to [1], we have introduced an additional test, viz., Test 3, for the case of non-Gaussian noise for reasons that will become apparent later. Note that if all sensors are honest, so that $\mathbf{z}[t] \equiv \mathbf{x}[t]$, then the reported measurements will pass all of the above tests almost surely.

Define

$$\mathbf{v}[t+1] \coloneqq \mathbf{z}[t+1] - A\mathbf{z}[t] - Bg_t(\mathbf{z}^t) - B\mathbf{e}[t] - \mathbf{w}[t+1]. \tag{7.6}$$

This can be regarded as the additive distortion introduced by the sensors to the process noise at time $t+1$. Note that if all sensors are honest, then the additive distortion $\mathbf{v}[t] \equiv 0$. The fundamental security guarantee of DW [1, 62, 63] establishes that under certain conditions on the system parameters, if the reported sequence of measurements $\{\mathbf{z}\}$ pass the aforementioned tests, then the additive distortion can only be of zero power, no matter what attack strategy the malicious sensors that may be present in the system employ. That is,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} ||\mathbf{v}[k]||^2 = 0. \tag{7.7}$$

This result has been established in [1, 62] in the context of Gaussian systems where the process noise, the measurement noise, as well as the watermark are i.i.d. Gaussian random processes, and it is the basis of other results which guarantee that the control performance of the system cannot be degraded by the malicious sensors.

In this chapter, we extend the above result to systems of the form (7.1) which are affected by noise having an arbitrary distribution P_W. Given a noise distribution P_W, unless the distribution P_E of the watermark is appropriately designed, the fundamental security guarantee of DW does not hold, as illustrated by the following example.

Example: Consider the scalar system

$$x[t+1] = 0.9x[t] + u[t] + w[t+1], \tag{7.8}$$

where $\{w\}$ is a Bernoulli random process which takes the value 0 with probability $\frac{1}{4}$ and the value 1 with probability $\frac{3}{4}$, and is i.i.d. across time.

Consider an output variance-minimizing stationary linear feedback policy $g_t(z^t) = -0.9z[t]$,

and a watermark that is drawn from a Bernoulli distribution i.i.d. across time, with $\mathbb{P}(e[t] = 0) = \frac{1}{2}$ and $\mathbb{P}(e[t] = 1) = \frac{1}{2}$ for all t. The closed loop system evolves as

$$x[t+1] = 0.9(x[t] - z[t]) + e[t] + w[t+1]. \tag{7.9}$$

The actuator subjects the reported sequence of measurements $\{z\}$ to Tests (7.3), (7.4) and (7.5), and declares the presence of malicious sensors if any of them fail. In what follows, we construct a specific attack that passes all three tests, but nevertheless successfully introduces an additive distortion of non-zero power.

Note first that from (7.9), it follows that $s[t+1] \coloneqq x[t+1] - 0.9x[t] + 0.9z[t] = e[t] + w[t+1]$, so that $s[t+1] \in \{0, 1, 2\}$ for every t. Now, consider the attack

$$z[t+1] = \begin{cases} 0.9 & \text{if } s[t+1] = 0, \\ 0.7 & \text{if } s[t+1] = 1, \\ 2.1 & \text{if } s[t+1] = 2. \end{cases} \tag{7.10}$$

Then, the left-hand side (LHS) of (7.3) becomes $\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} z^2[k+1]$. Since $z[k+1]$ is distributed i.i.d. across time and takes the value 0.9 with probability $\frac{1}{8}$, the value 0.7 with probability $\frac{1}{2}$, and the value 2.1 with probability $\frac{3}{8}$, we have that the LHS of (7.3) almost surely equals $\mathbb{E}(z^2[t+1]) = 2$, which equals the right-hand side (RHS) of (7.3). It is equally straightforward to verify that the other two Tests (7.4) and (7.5) are also passed by the sequence $\{z\}$. Since this attack passes all three tests, it will go undetected by the attack detector. However, the additive distortion that this attack introduces has power $\frac{3}{20} \neq 0$. Hence, the watermark distribution that has been chosen fails to provide the fundamental security guarantee (7.7).

On the other hand, if the watermark distribution P_E is chosen to be Bernoulli which takes the value 0 with probability $\frac{1}{4}$, the value 1 with probability $\frac{3}{4}$, and i.i.d. across time, then, as established in Theorem 7 of [1], the fundamental security guarantee (7.7) holds, and no attack that introduces any non-zero power additive distortion can manage to pass all three tests. □

As the above example illustrates, given a linear stochastic system, the probability distribution of the watermark has to be carefully designed taking into account the noise probability distribution if the fundamental security guarantee of Dynamic Watermarking is to hold. This brings us to the central question addressed in this chapter: Given a closed loop system of the form (7.2), and the probability distribution P_W of the process noise, what are the necessary and sufficient conditions that the probability distribution P_E of the watermark should satisfy so that the fundamental security guarantee (7.7) holds whenever the reported measurements $\{\mathbf{z}\}$ pass Tests (7.3), (7.4) and (7.5)? The next section answers this question.

7.3 Design of Security-Guaranteeing Dynamic Watermarks

In this section, we derive the necessary and sufficient conditions that the probability distribution of the watermark should satisfy in (7.2) in order to provide the fundamental security guarantee. We do this in two steps. First, in order to expose clearly the main ideas involved in studying Dynamic Watermarking in a setting where the noise distribution is arbitrary, we simplify the problem and formulate a core problem so that the only complications that arise in its analysis are those that are introduced by the arbitrariness of the process noise distribution. Consequently, we first restrict attention to system (7.2) where the matrix $A = 0$, the control policy is identically equal to zero, and the adversary is restricted to employing stationary Markov attack policies. Here, by a stationary Markov attack policy, we are employing the terminology of Markov Decision Processes to refer to an attack policy which chooses the measurements to be reported at any time t as a function of only the state of the system at that time. Then we show how these conditions also extend to the more general case where the matrix A is arbitrary, the control policy is arbitrary, and the adversary employs an arbitrary, possibly history-dependent and randomized attack policy which is unknown to the honest nodes in the system. We begin with the first of these cases.

7.3.1 The core problem with zero dynamics matrix, zero control policy, and stationary Markov attack policies

We consider in this subsection a special case of system (7.2) with $A = 0$ and $g_t = 0$. In this case, the system (7.2) reduces to

$$\mathbf{x}[t+1] = B\mathbf{e}[t] + \mathbf{w}[t+1], \tag{7.11}$$

and the additive distortion $\mathbf{v}[t+1]$ defined in (7.6), reduces to

$$\mathbf{v}[t+1] = \mathbf{z}[t+1] - B\mathbf{e}[t] - \mathbf{w}[t+1]. \tag{7.12}$$

Define functions $\mathbf{\Psi}^W$ and $\mathbf{\Psi}^E$ by

$$\mathbf{\Psi}^W(\mathbf{x}[t+1]) \coloneqq \mathbb{E}[\mathbf{w}[t+1]\big|\mathbf{x}[t+1]], \tag{7.13}$$

and

$$\mathbf{\Psi}^E(\mathbf{x}[t+1]) \coloneqq \mathbb{E}[\mathbf{e}[t]\big|\mathbf{x}[t+1]]. \tag{7.14}$$

Note that the joint probability distribution of $\mathbf{\Psi}^W(\mathbf{x}[t+1])$ and $\mathbf{\Psi}^E(\mathbf{x}[t+1])$ does not depend on the time $t+1$, and hence we will henceforth omit the argument $\mathbf{x}[t+1]$ in the following, and simply use $\mathbf{\Psi}^W$ and $\mathbf{\Psi}^E$ to denote the above random vectors, noting that the results hold for all $\mathbf{x}[t+1]$.

Recall from Section 4.1 that the notation Ψ_i^W denotes the i^{th} component of the vector $\mathbf{\Psi}^W$. The following theorem provides the necessary and sufficient conditions on the watermark to ensure that the additive distortion can only be of zero power.

Theorem 14. *Consider the system (7.11). Suppose that the adversarial sensors employ a stationary attack policy, and that the reported sequence of measurements $\{\mathbf{z}\}$ passes the tests (7.3), (7.4)*

and (7.5). If

$$\Psi_i^W \in \mathit{Span}\{\Psi_1^E, \ldots, \Psi_m^E, 1\} \tag{7.15}$$

for all $i \in \{1, \ldots, p\}$*, then,*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} ||\mathbf{v}[k]||^2 = 0. \tag{7.16}$$

Moreover, if (7.15) does not hold, then, there exists a stationary attack passing tests (7.3), (7.4) and (7.5), but for which (7.16) does not hold.

Proof. Using (7.12) and (7.11), we have that

$$\mathbf{v}[t+1] = \mathbf{z}[t+1] - \mathbf{x}[t+1]. \tag{7.17}$$

Since the attack policy employed by the adversary is assumed to be a stationary Markov policy, the measurements reported by the sensors at time $t+1$ are a function of only their observations at time $t+1$, and therefore, $\mathbf{z}[t+1] \in \sigma(\mathbf{x}[t+1])$, the $\sigma-$algebra generated by $\mathbf{x}[t+1]$. Consequently,

$$\mathbf{v}[t+1] \in \sigma(\mathbf{x}[t+1]). \tag{7.18}$$

The attack policy being a stationary Markov policy, $\{\mathbf{z}\}$ is a sequence of i.i.d. random vectors, and therefore, the LHS of (7.4) almost surely equals $\mathbb{E}[\mathbf{e}[0](\mathbf{z}[1] - B\mathbf{e}[0])^T]$. From (7.17), (7.11), and (7.4),

$$\mathbb{E}(\mathbf{e}[0]\mathbf{v}^T[1]) = 0. \tag{7.19}$$

Also, the LHS of (7.3) almost surely equals $\mathbb{E}(\mathbf{z}[1]\mathbf{z}^T[1])$. Using (7.17), (7.3), (7.11) and (7.19),

and carrying out some algebra, yields

$$\mathbb{E}(\mathbf{v}[1]\mathbf{v}^T[1]) + \mathbb{E}(\mathbf{w}[1]\mathbf{v}^T[1]) + \mathbb{E}(\mathbf{v}[1]\mathbf{w}^T[1]) = 0. \tag{7.20}$$

Similarly, using (7.17) and (7.5) gives

$$\mathbb{E}(\mathbf{v}[1]) = 0. \tag{7.21}$$

Now, note that

$$\mathbb{E}(\mathbf{e}[0]\mathbf{v}^T[1]) = \mathbb{E}\,\mathbb{E}\{\mathbf{e}[0]\mathbf{v}^T[1]\big|\mathbf{x}[1]\} = \mathbb{E}(\mathbf{\Psi}^E(\mathbf{x}[1])\mathbf{v}^T[1]), \tag{7.22}$$

where the last equality follows from (7.18). Using this with (7.19) gives

$$\mathbb{E}(\mathbf{\Psi}^E(\mathbf{x}[1])\mathbf{v}^T[1]) = 0. \tag{7.23}$$

Combining (7.21), (7.23) and (7.15) gives

$$\mathbb{E}(\mathbf{\Psi}^W(\mathbf{x}[1])\mathbf{v}^T[1]) = 0. \tag{7.24}$$

Now,

$$\mathbb{E}(\mathbf{\Psi}^W(\mathbf{x}[1])\mathbf{v}^T[1]) = \mathbb{E}\big\{\mathbb{E}\big[\mathbf{w}[1]\big|\mathbf{x}[1]\big]\mathbf{v}^T[1]\big\} = \mathbb{E}(\mathbf{w}[1]\mathbf{v}^T[1]), \tag{7.25}$$

where the second equality follows from (7.18) and the law of iterated expectations. Combining this with (7.24) gives $\mathbb{E}(\mathbf{w}[1]\mathbf{v}^T[1]) = 0$, and substituting this in (7.20) establishes (7.16).

Now we turn to the next assertion. Suppose that (7.15) does not hold. Define

$$\mathcal{S} \coloneqq \mathit{Span}\{\Psi_1^E, \ldots, \Psi_m^E, 1\},$$

and for each $i \in \{1, \ldots, p\}$, $\Phi_i \coloneqq \Psi_i^W - P_{\mathcal{S}}(\Psi_i^W)$, where $P_{\mathcal{S}}$ denotes the operator that projects a random variable onto the subspace $\mathcal{S}$. It follows that for all $i \in \{1, \ldots, p\}$,

$$\mathbb{E}(\Phi_i \Psi_j^E) = 0 \tag{7.26}$$

for all $j \in \{1, \ldots, m\}$, and

$$\mathbb{E}\Phi_i = 0. \tag{7.27}$$

Now, define $\mathbf{\Phi} \coloneqq [\Phi_1 \ \ldots \ \Phi_p]^T$ and consider the attack $\mathbf{z}[t+1] = \mathbf{x}[t+1] - 2\mathbf{\Phi}(\mathbf{x}[t+1])$. The attacker can compute $\mathbf{z}[t+1]$ since it depends only on $\mathbf{x}[t+1]$, which it can measure. Then, $\mathbf{v}[t+1] = -2\mathbf{\Phi}(\mathbf{x}[t+1])$. It follows from (7.27) that $\mathbb{E}(\mathbf{v}[t+1]) = 0$, and from (7.26) that $\mathbb{E}(\mathbf{\Psi}^E(\mathbf{x}[t+1])\mathbf{v}^T[t+1]) = 0$. Hence, from (7.22), it follows that $\mathbb{E}(\mathbf{e}[t]\mathbf{v}^T[t+1]) = 0$. Hence, the attack passes Tests (7.4) and (7.5). Also, for all $i, j \in \{1, \ldots, p\}$,

$$\begin{aligned}
\mathbb{E}(v_i v_j) + \mathbb{E}(v_i \Psi_j^W) + \mathbb{E}(\Psi_i^W v_j) &= 4\mathbb{E}(\Phi_i \Phi_j) - 2\mathbb{E}(\Phi_i \Psi_j^W) - 2\mathbb{E}(\Psi_i^W \Phi_j) \\
&= 4\mathbb{E}(\Phi_i \Phi_j) - 2\mathbb{E}[\Phi_i(\Phi_j + P_{\mathcal{S}}(\Psi_j^W))] \\
&\quad - 2\mathbb{E}[(\Phi_i + P_{\mathcal{S}}(\Psi_i^W))\Phi_j] = 0,
\end{aligned}$$

where the last equality follows using (7.26) and (7.27) since $P_{\mathcal{S}}(\Psi_i^W)$ and $P_{\mathcal{S}}(\Psi_j^W)$, belonging to the subspace spanned by $\{\Psi_1^E, \ldots, \Psi_m^E, 1\}$, are some linear combinations of these random variables. Combining the above equality with (7.25) implies that the attack satisfies (7.20), which in turn implies that the attack also passes test (7.3). Hence, the attack passes all three tests. However, since (7.15) does not hold, $\mathbf{\Phi} \neq 0$, and therefore $\mathbb{E}(||\mathbf{v}[t+1]||^2) \neq 0$, implying that the fundamental security guarantee (7.16) does not hold. □

The intuition behind condition (7.15) is as follows. The condition essentially implies that the adversary cannot “separate" the process noise and the watermark along any subspace of the system’s state space. Now, note that in order for the adversary to pass Test 1, it has to introduce a

distortion that is correlated with the observations $B\mathbf{e}[t] + \mathbf{w}[t+1]$ in (7.11). However, since the adversary cannot separate the noise and the watermark perfectly, the distortion that it introduces will have to be correlated with both process noise as well as the watermark. Test 2, however, ensures that any distortion introduced is not correlated with the watermark. Consequently, the adversary is constrained to introducing a distortion that can only be of zero power.

Having characterized security-guaranteeing watermarks in a simplified system context, we now address the general case where the matrix A in (7.1) is arbitrary, the control policy is arbitrary, and the adversary is allowed to implement any arbitrary, history-dependent attack policy. We show that in this general system context, a condition analogous to (7.15) serves as the necessary and sufficient condition on the watermark distribution in order for the fundamental security guarantee to hold.

7.3.2 The case of general MIMO systems with arbitrary control policy and arbitrary history-dependent attack policies

Consider the closed loop system (7.2), an arbitrary subset of whose sensors may be malicious, and suppose that the attack detector conducts the tests (7.3), (7.4) and (7.5). Define

$$\mathbf{\Psi}^W(B\mathbf{e}[t] + \mathbf{w}[t+1]) \coloneqq \mathbb{E}(\mathbf{w}[t+1] \big| B\mathbf{e}[t] + \mathbf{w}[t+1]), \tag{7.28}$$

and

$$\mathbf{\Psi}^E(B\mathbf{e}[t] + \mathbf{w}[t+1]) \coloneqq \mathbb{E}(\mathbf{e}[t] \big| B\mathbf{e}[t] + \mathbf{w}[t+1]). \tag{7.29}$$

In the above definitions, we have reused notation from the previous subsection owing to the fact that both (7.13) and (7.28) signify analogous quantities, viz., the Minimum Mean-Squared Error (MMSE) estimate of the process noise computable by the malicious sensors in the two system models, and similarly, both (7.14) and (7.29) denote the MMSE estimate of the watermark.

The following theorem establishes in the context of a fully observed MIMO system the neces-

sary and sufficient conditions that the watermark distribution should satisfy in order for the fundamental security guarantee to hold, and is a central result of this chapter.

Theorem 15. *Consider the system (7.2) and suppose that the measurements* $\{\mathbf{z}\}$ *reported by the sensors pass Tests (7.3), (7.4) and (7.5). If*

$$\Psi_i^W \in \mathit{Span}\{\Psi_1^E, \ldots, \Psi_m^E, 1\} \tag{7.30}$$

for all $i \in \{1, \ldots, p\}$*, then,*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} ||\mathbf{v}[k]||^2 = 0. \tag{7.31}$$

Moreover, if (7.30) does not hold, then there exists an attack passing Tests (7.3), (7.4) and (7.5), but for which (7.31) does not hold.

Proof. Since the reported measurements pass Test (7.4), it follows from (7.6) that

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} \mathbf{e}[k]\mathbf{v}^T[k+1] = 0. \tag{7.32}$$

Define $\widetilde{\mathbf{e}}[k] \coloneqq \mathbf{e}[k] - \mathbf{\Psi}^E[k]$. Then, the above equality becomes

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (\mathbf{\Psi}^E[k] + \widetilde{\mathbf{e}}[k])\mathbf{v}^T[k+1] = 0.$$

Now, the reported measurement $\mathbf{z}[k+1]$ at time $k+1$ can be a function of the past and the current observations of the sensors. In addition, a malicious sensor at each time $k+1$ could also adapt its reported measurement to a random variable $\theta[k+1]$ that it generates, which is independent of all random variables that have been realized until that time. Define $\mathcal{G}_k \coloneqq \sigma(\mathbf{x}^k, \theta^k, \mathbf{e}^{k-2})$. It follows that $\mathbf{z}[k+1] \in \sigma(\mathbf{x}^{k+1}, \theta^{k+1}) \subseteq \mathcal{G}_{k+1}$, which in turn implies that $\mathbf{v}[k+1] \in \mathcal{G}_{k+1}$. It is also straightforward to verify that $(\widetilde{\mathbf{e}}[k-1], \mathcal{G}_{k+1})$ is a Martingale difference sequence. Using the

Martingale Stability Theorem [49], we have that for all $i \in \{1, \ldots, m\}$ and all $j \in \{1, \ldots, p\}$,

$$\sum_{k=0}^{T-1} \widetilde{e}_i[k] v_j[k+1] = o(\sum_{k=0}^{T-1} v_j^2[k+1]) + O(1).$$

Substituting this in the previous equality gives

$$\sum_{k=0}^{T-1} \mathbf{\Psi}^E[k] \mathbf{v}^T[k+1] = \begin{bmatrix} o(\sum_{k=1}^{T} v_1^2[k]) & \cdots & o(\sum_{k=1}^{T} v_p^2[k]) \\ \vdots & \ddots & \vdots \\ o(\sum_{k=1}^{T} v_1^2[k]) & \cdots & o(\sum_{k=1}^{T} v_p^2[k]) \end{bmatrix} + o(T). \tag{7.33}$$

Since the reported measurements also pass Test (7.3), we have that

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} B\mathbf{e}[k]\mathbf{v}^T[k+1] + \mathbf{w}[k+1]\mathbf{v}^T[k+1] + \mathbf{v}[k+1]\mathbf{e}^T[k]B^T + \mathbf{v}[k+1]\mathbf{w}^T[k+1] + \mathbf{v}[k+1]\mathbf{v}^T[k+1] = 0.$$

Substituting (7.32) in the above equality yields

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} \mathbf{w}[k+1]\mathbf{v}^T[k+1] + \mathbf{v}[k+1]\mathbf{w}^T[k+1] + \mathbf{v}[k+1]\mathbf{v}^T[k+1] = 0. \tag{7.34}$$

Similarly, it follows from (7.5) that

$$\sum_{k=0}^{T-1} \mathbf{v}[k+1] = o(T). \tag{7.35}$$

Define $\widetilde{\mathbf{w}}[t+1] \coloneqq \mathbf{w}[t+1] - \mathbf{\Psi}^W[t+1]$. Then, $(\widetilde{\mathbf{w}}[k], \mathcal{G}_{k+1})$ is a martingale difference sequence, and using the Martingale Stability Theorem, we have that for all $i, j \in \{1, \ldots, p\}$,

$$\sum_{k=0}^{T-1} \widetilde{w}_i[k+1] v_j[k+1] = o(\sum_{k=0}^{T-1} v_j^2[k+1]) + O(1). \tag{7.36}$$

We also have using (7.30), (7.33), and (7.35) that for all $i, j \in \{1, \ldots, p\}$,

$$\sum_{k=0}^{T-1} \Psi_i^W[k+1] v_j[k+1] = o(\sum_{k=0}^{T-1} v_j^2[k+1]) + o(T). \tag{7.37}$$

Adding (7.36) and (7.37) gives

$$\sum_{k=0}^{T-1} \mathbf{w}[k+1]\mathbf{v}^T[k+1] = \begin{bmatrix} o(\sum_{k=1}^{T} v_1^2[k]) & \cdots & o(\sum_{k=1}^{T} v_p^2[k]) \\ \vdots & \ddots & \vdots \\ o(\sum_{k=1}^{T} v_1^2[k]) & \cdots & o(\sum_{k=1}^{T} v_p^2[k]) \end{bmatrix} + o(T). \tag{7.38}$$

Substituting (7.38) in (7.34) establishes (7.31).

We now turn to the next assertion. Suppose that (7.30) does not hold. As in the proof of Theorem 1, define $\mathcal{S} \coloneqq \mathit{Span}\{\Psi_1^E, \ldots, \Psi_m^E, 1\}$, and define for each $i \in \{1, \ldots, p\}$, $\Phi_i \coloneqq \Psi_i^W - P_{\mathcal{S}}(\Psi_i^W)$, where $P_{\mathcal{S}}$ is the operator that projects a random variable on the subspace $\mathcal{S}$. Let $\mathbf{\Phi} \coloneqq [\Phi_1 \ldots \Phi_p]^T$ and consider the attack $\mathbf{z}[t+1] = A\mathbf{z}[t] + (\mathbf{x}[t+1] - A\mathbf{x}[t]) - 2\mathbf{\Phi}[t+1]$, so that $\mathbf{v}[t+1] = -2\mathbf{\Phi}[t+1]$. Following similar arguments as in the proof of Theorem 1, it can be verified that this attack passes all three tests (7.3), (7.4) and (7.5), but introduces non-zero powered additive distortion to the process noise. □

7.4 Concluding Remarks

The approach of Dynamic Watermarking provides provable security guarantees for cyberphysical systems from arbitrary sensor attacks. Prior works have developed the theory of Dynamic Watermarking and have established the fundamental security guarantee that it provides in the context of systems affected by process and measurement noises that are Gaussian random processes. In this chapter, we have gone beyond the class of Gaussian systems and have considered finite-dimensional linear systems affected by arbitrarily distributed process noise. We have shown how the fundamental security guarantee of Dynamic Watermarking fails in these system contexts when the watermark distribution is not appropriately chosen. Consequently, we have derived for the class of perfectly observed, finite-dimensional linear systems with arbitrarily process noise distribution,

the necessary and sufficient conditions that the distribution of the watermark should satisfy in order for it to provide the fundamental security guarantee. Extensions of this result to the case of partially observed systems and systems affected by colored noise process could potentially constitute the foundations of a theory for security-guaranteeing watermark design.

8. ON THE WATERMARK-SECURABLE SUBSPACE OF A LINEAR SYSTEM

8.1 Introduction

In prior chapters, the approach of Dynamic Watermarking was established as providing provable security guarantees against arbitrary sensor attacks in a variety of control system settings. Specifically, it was shown that in a stochastic system in which Dynamic Watermarking is employed, under appropriate conditions, any adversarial sensor that may be present in the system cannot distort the innovations process of the output by more than a zero-power signal if it wishes to remain undetected. In Chapters 3, 4 and 5, the notions of securable and unsecurable subspaces of a linear system were introduced and shown to have important operational meanings in the context of secure control. Specifically, it was shown that in the absence of Dynamic Watermarking, the state space can be decomposed into two orthogonal subspaces, called the securable and the unsecurable subspaces, such that the malicious sensors and actuators cannot degrade the state estimation performance of the honest sensors and actuators along the securable subspace of the system if they wish to remain undetected. In this chapter, we consider a system wherein Dynamic Watermarking is employed and an arbitrary subset of both sensors and actuators could be malicious. Analogous to prior results [48, 50], the state space of a system employing Dynamic Watermarking too can be decomposed into two orthogonal subspaces, called the watermark-securable and the watermark-unsecurable subspaces, such that the malicious sensors and actuators cannot degrade the state estimation performance of the honest sensors and actuators along the watermark-securable subspace if they are to remain undetected. The watermark-securable subspace is larger than the securable subspace that can be guaranteed in the absence of Dynamic Watermarking. In this chapter, we characterize a certain subset of the watermark-securable subspace.

The rest of the chapter is organized as follows. Section 8.2 describes the problem setup. Section 8.3 contains certain definitions and results that will be used to establish the main result of the chapter. Section 8.4 presents the main result. Section 8.5 concludes the chapter.

Notation: Given a matrix B, we denote by $\mathcal{R}(B)$ the range space of B, and by $\mathcal{N}(B)$ the null space of B. We denote by I an identity matrix whose dimensions will be clear from the context. Given a vector $\mathbf{x}$ and a subspace $\mathcal{S}$, we denote by $\mathbf{x}_{\mathcal{S}}$ the projection of the vector $\mathbf{x}$ on the subspace $\mathcal{S}$. Given two subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$, we denote by $\mathcal{S}_1 \oplus \mathcal{S}_2$ their span, and by $\mathcal{S}_{1\mathcal{S}_2}$ the projection of $\mathcal{S}_1$ on $\mathcal{S}_2$. The letter P always denotes a projection matrix; by $P_{\mathcal{X}}$, we denote a matrix which projects a vector right-multiplying it onto the subspace $\mathcal{X}$.

8.2 Problem Formulation

Consider a p^{th} order perfectly-observed linear stochastic system described by

$$\begin{aligned} \mathbf{x}[t+1] &= A\mathbf{x}[t] + B\mathbf{u}[t] + \mathbf{w}[t+1], \\ \mathbf{y}[t+1] &= C\mathbf{x}[t+1], \end{aligned} \tag{8.1}$$

where $\mathbf{x}[t] \in \mathbb{R}^p$ is the state of the system at time t, $\mathbf{u}[t] \in \mathbb{R}^m$ is the input applied to the system at time t, $\mathbf{w}[t+1] \sim \mathcal{N}(0, \sigma_W^2 I)$ is the process noise affecting the system at time $t+1$, and $\mathbf{y}[t] \in \mathbb{R}^n$ is the system output at time t. In this chapter, we restrict attention to perfectly-observed systems, i.e., $C = I$. The process noise is independent and identically distributed (iid) across time.

As mentioned in Section 8.1, we consider a setting wherein an arbitrary subset of sensors and actuators in system (8.1) could be malicious. As always, we allow for the malicious nodes in the system to know the identities of all other malicious nodes, and consequently for the possibility that they exchange information among themselves via an underlying communication network, to carry out coordinated attack strategies. The honest nodes do not know the identities of the honest or the malicious nodes, or even if there are any malicious nodes in the system.

At each time t, each sensor is supposed to report the measurement that it observes to all other sensors and actuators in the system. To allow for greater power for the malicious nodes, we suppose that the honest sensors first announce their measurements to all other nodes in the system, after which the malicious sensors report their measurements. This sequence of reporting allows the malicious sensors to adapt the measurements that they report at each time to the measurements of

all honest sensors until that time. In addition, they could also adapt their reported measurements to the inputs of the malicious actuators, including their inputs at future times, if they wish to do so.

We now turn to the protocols followed by the actuators. As before, only the honest actuators follow this protocol, while the malicious actuators can behave arbitrarily. For each $i \in \{1, \ldots, m\}$, we denote by $\{g_0^i, g_1^i, \ldots\}$ an arbitrary and possibly history-dependent control policy that actuator i is supposed to employ in determining its inputs. The control policy-specified input $u_i^g[t]$ of actuator i at time t is

$$u_i^g[t] = g_t^i(\mathbf{z}^t). \tag{8.2}$$

The honest actuators watermark their control inputs. The net input $u_i[t]$ applied by an honest actuator i at time t is therefore

$$u_i[t] = g_t^i(\mathbf{z}^t) + e_i[t].$$

The malicious actuators can choose the inputs that they apply in an arbitrary manner. For a malicious actuator j, denote by $u_j[t]$ the input that it applies at time t, and let $e_j[t] \coloneqq u_j[t] - g_t^j(\mathbf{z}^t)$. Since actuator j is not honest, the statistics of the sequence $\{e_j\}$ may not be equal to the statistics of the watermark.

Let $\mathbf{e}[t] \coloneqq [e_1[t] \ \cdots \ e_m[t]]^T$. Denote by $\mathbf{e}^H[t]$ the sub-vector of $\mathbf{e}[t]$ corresponding to the entries of the honest actuators, and by $\mathbf{e}^M[t]$ the sub-vector of $\mathbf{e}[t]$ corresponding to the entries of the malicious actuators. Similarly, denote by B^H the sub-matrix of B consisting of those columns of B corresponding to the honest actuators, and by B^M the sub-matrix of B consisting of those columns of B corresponding to the malicious actuators. Then, system (8.1) evolves in closed loop as

$$\begin{aligned} \mathbf{x}[t+1] = A\mathbf{x}[t] + B\mathbf{g}_t(\mathbf{z}^t) + B^H\mathbf{e}^H[t] + B^M\mathbf{e}^M[t] + \mathbf{w}[t+1], \\ \mathbf{y}[t+1] = C\mathbf{x}[t+1], \end{aligned} \tag{8.3}$$

where $\mathbf{g}_t(\mathbf{z}^t) \coloneqq [g_t^1(\mathbf{z}^t) \ \cdots \ g_t^m(\mathbf{z}^t)]^T$, and $\mathbf{e}^H[t] \sim \mathcal{N}(0, \sigma_e^2 I)$ and is i.i.d. across time. We also denote by C^H the submatrix of C (which in this chapter is the identity matrix) formed by its rows that correspond to the honest sensors.

In order to check whether or not there are any malicious nodes in the system, the following tests are carried out.

Test 1: Each honest node checks if

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (\mathbf{z}[k+1] - A\mathbf{z}[k] - B\mathbf{g}_k(\mathbf{z}^k))^{TT} = BB^T\sigma_e^2 + \sigma_w^2 I. \tag{8.4}$$

Test 2: Each honest actuator i checks if

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} e_i[k](\mathbf{z}[k+1] - A\mathbf{z}[k]) = B_{\cdot i}\sigma_e^2. \tag{8.5}$$

Our goal is to characterize the extent to which the malicious sensors and actuators can distort the state estimate of the honest actuators while remaining undetected by Tests (8.4) and (8.5). We show that along a certain subspace of the state space $\mathbb{R}^p$ of system (8.3), no matter what attack strategy the malicious nodes employ, they cannot degrade the "state estimation performance" of the honest actuators if they are to remain undetected by Tests (8.4) and (8.5). This subspace is characterized in the following sections.

8.3 Towards the Watermark-Securable and the Watermark-Unsecurable Subspaces of a Linear System

In this section, we present a few preparatory results for establishing the main result of the chapter. Recall the definitions of the unsecurable subspace $\mathcal{V}$, the securable subspace $\mathcal{S}$, the $k-$step unsecurable subspaces $\mathcal{U}_k$, and the $k-$step securable subspaces $\mathcal{S}_k$ from Chapter 4.

Lemma 8. *The securable subspace $\mathcal{S}$ is the span of the $k-$step securable subspaces, i.e.,*

$$\mathcal{S} = \mathcal{S}_0 \oplus \ldots \oplus \mathcal{S}_{r-1}.$$

Proof. See Lemma 2 of [50]. □

Lemma 9. *Let* $k \in \{0, \ldots, r-1\}$. *Then,*

$$\mathcal{U}_k = [\oplus_{i=k}^{r-1} \mathcal{S}_k] \oplus \mathcal{V}. \tag{8.6}$$

Proof. Since $\mathcal{U} = \mathcal{U}_r \subset \mathcal{U}_{r-1} \subset \ldots \subset \mathcal{U}_1 \subset \mathcal{U}_0 \coloneqq \mathbb{R}^p$, it follows that there exist sets $\mathcal{B}_0, \ldots, \mathcal{B}_r$ of orthonormal basis vectors for the subspaces $\mathcal{U}_0, \ldots, \mathcal{U}_r$ respectively such that $\mathcal{B}_r \subset \ldots \subset \mathcal{B}_0$. Then,

$$[\oplus_{i=k}^{r-1} \mathcal{S}_i] \oplus \mathcal{V} = [\oplus_{i=k}^{r-1} \mathit{Span}\{\mathcal{B}_i \setminus \mathcal{B}_{i+1}\}] \oplus \mathcal{V} = [\mathit{Span}\{\mathcal{B}_k \setminus \mathcal{B}_r\}] \oplus \mathcal{V} = \mathit{Span}\{\mathcal{B}_k\} = \mathcal{U}_k. \tag{8.7}$$

□

Lemma 10. *Let* $k \in \{1, \ldots, r-1\}$ *and* $\mathbf{s} \in \mathcal{S}_k$. *There exists* $\mathbf{d}$ *such that*

$$(A\mathbf{s} + B^M \mathbf{d}) \in [\oplus_{i=k-1}^{r-1} \mathcal{S}_i] \oplus \mathcal{V}. \tag{8.8}$$

Proof. The result holds trivially for $k = 1$ since $[\oplus_{i=0}^{r-1} \mathcal{S}_i] \oplus \mathcal{V} = \mathbb{R}^p$, the entire state space. Consider, therefore, the case when $k > 1$. It follows from the definition of $\mathcal{S}_k$ that $\mathcal{S}_k \subset \mathcal{U}_k$, and therefore, $\mathbf{s} \in \mathcal{U}_k$. It follows from the definition of $\mathcal{U}_k$ that there exists $\mathbf{d}$ such that $A\mathbf{s} + B^M \mathbf{d} \in \mathcal{V}_{k-1}$. Combining this with (8.6) yields the desired result. □

Lemma 11. *Let* $k \in \{1, \ldots, r-1\}$ *and* $\mathbf{s} \in \mathcal{S}_k \setminus \{0\}$. *For every* $\mathbf{d}$,

$$(A\mathbf{s} + B^M \mathbf{d})_{\oplus_{i=0}^{k-1} \mathcal{S}_i} \neq 0. \tag{8.9}$$

Proof. Suppose for contradiction that (8.9) does not hold. Then, there exists $\mathbf{d}$ such that $(A\mathbf{s} + B^M \mathbf{d})_{\oplus_{i=0}^{k-1} \mathcal{S}_i} = 0$, implying that $(A\mathbf{s} + B^M \mathbf{d}) \in [\oplus_{i=k}^{r-1} \mathcal{S}_i] \oplus \mathcal{V}$. Combining this with (8.6), we have that $A\mathbf{s} + B^M \mathbf{d} \in \mathcal{U}_k$. Therefore, we have that $\mathbf{s} \in \mathcal{S}_k$ and that that there exists $\mathbf{d}$ such

that $A\mathbf{s} + B^M\mathbf{d} \in \mathcal{U}_k$ implying that $\mathbf{s} \in \mathcal{V}_{k+1}$. Since $\mathcal{S}_k \cap \mathcal{V}_{k+1} = \{0\}$, it follows that $\mathbf{s} = 0$, a contradiction. □

Lemma 12. *Let* $k \in \{1, \ldots, r-1\}$ *and* $\mathbf{s} \in \mathcal{S}_k \setminus \{0\}$. *For every* $\mathbf{d}$,

$$(A\mathbf{s} + B^M\mathbf{d})_{[\oplus_{i=0}^{k-1}\mathcal{S}_i]\cap([\mathcal{R}(B^M)]_{\oplus_{i=0}^{k-1}\mathcal{S}_i})^\perp} \neq 0. \tag{8.10}$$

Proof. Note first that for any two subspaces $\mathcal{C}_1, \mathcal{C}_2$ such that $\mathcal{C}_1 \subseteq \mathcal{C}_2$, we have that $\mathcal{C}_2 = \mathcal{C}_1 \oplus [\mathcal{C}_1^\perp \cap \mathcal{C}_2]$. Substituting $\mathcal{C}_2 = \oplus_{i=0}^{k-1}\mathcal{S}_i$ and $\mathcal{C}_1 = [\mathcal{R}(B^M)]_{\oplus_{i=0}^{k-1}\mathcal{S}_i}$, we have

$$\oplus_{i=0}^{k-1}\mathcal{S}_i = [\mathcal{R}(B^M)]_{\oplus_{i=0}^{k-1}\mathcal{S}_i} \oplus \left[\left([\mathcal{R}(B^M)]_{\oplus_{i=0}^{k-1}\mathcal{S}_i}\right)^\perp \cap [\oplus_{i=0}^{k-1}\mathcal{S}_i]\right].$$

Now, suppose for contradiction that (8.10) does not hold. Then, there exists $\mathbf{d}$ such that

$$\begin{aligned}(A\mathbf{s} + B^M\mathbf{d})_{\oplus_{i=0}^{k-1}\mathcal{S}_i} &= (A\mathbf{s} + B^M\mathbf{d})_{([\mathcal{R}(B^M)]_{\oplus_{i=0}^{k-1}\mathcal{S}_i})^\perp\cap[\oplus_{i=0}^{k-1}\mathcal{S}_i]} \oplus (A\mathbf{s} + B^M\mathbf{d})_{[\mathcal{R}(B^M)]_{\oplus_{i=0}^{k-1}\mathcal{S}_i}} \\ &= (A\mathbf{s} + B^M\mathbf{d})_{[\mathcal{R}(B^M)]_{\oplus_{i=0}^{k-1}\mathcal{S}_i}},\end{aligned}$$

where the last equality follows from the assumption that (8.10) does not hold. Since $(A\mathbf{s} + B^M\mathbf{d})_{[\mathcal{R}(B^M)]_{\oplus_{i=0}^{k-1}\mathcal{S}_i}} \in [\mathcal{R}(B^M)]_{\oplus_{i=0}^{k-1}\mathcal{S}_i}$, there exists $\bar{\mathbf{d}}$ such that

$$(B^M\bar{\mathbf{d}})_{\oplus_{i=0}^{k-1}\mathcal{S}_i} = -(A\mathbf{s} + B^M\mathbf{d})_{[\mathcal{R}(B^M)]_{\oplus_{i=0}^{k-1}\mathcal{S}_i}}.$$

Therefore,

$$(A\mathbf{s} + B^M(\mathbf{d} + \bar{\mathbf{d}}))_{\oplus_{i=0}^{k-1}\mathcal{S}_i} = (A\mathbf{s} + B^M\mathbf{d})_{\oplus_{i=0}^{k-1}\mathcal{S}_i} + (B^M\bar{\mathbf{d}})_{\oplus_{i=0}^{k-1}\mathcal{S}_i} = 0,$$

where the last equality follows from the preceding two equalities. This, however, contradicts Lemma 11. □

Lemma 13. *Let $k \in \{1, \ldots, r-1\}$. There exists $\epsilon > 0$ such that for all $\mathbf{s} \in \mathcal{S}_k$,*

$$\left|\left|(A\mathbf{s})_{[\oplus_{j=0}^{k-1}\mathcal{S}_j]\cap([\mathcal{R}(B^M)]_{\oplus_{j=0}^{k-1}\mathcal{S}_j})^\perp}\right|\right|^2 \geq \epsilon \, ||\mathbf{s}||^2. \tag{8.11}$$

Proof. Define $J(\mathbf{x}) \coloneqq ||P_{[\oplus_{j=0}^{k-1}\mathcal{S}_j]\cap([\mathcal{R}(B^M)]_{\oplus_{j=0}^{k-1}\mathcal{S}_j})^\perp} A\mathbf{x}||$. Let $\bar{\epsilon} \coloneqq \inf_{\{\bar{\mathbf{s}}\in\mathcal{S}_k,\, ||\bar{\mathbf{s}}||=1\}} J(\bar{\mathbf{s}})$. Let $\mathbf{s}^*$ attain the minimum of the continuous function J over the specified compact set. By (8.10), there exists a $\mathbf{d}$ such that $(A\mathbf{s}^*+B^M\mathbf{d})_{[\oplus_{i=0}^{k-1}\mathcal{S}_i]\cap([\mathcal{R}(B^M)]_{\oplus_{i=0}^{k-1}\mathcal{S}_i})^\perp} \neq 0$. Since $[\mathcal{R}(B^M)]_{[\oplus_{j=0}^{k-1}\mathcal{S}_j]\cap[\mathcal{R}(B^M)]^\perp_{\oplus_{j=0}^{k-1}\mathcal{S}_j}} = 0$, we have $[A\mathbf{s}^*]_{[\oplus_{i=0}^{k-1}\mathcal{S}_i]\cap([\mathcal{R}(B^M)]_{\oplus_{i=0}^{k-1}\mathcal{S}_i})^\perp} \neq 0$. Hence, $\bar{\epsilon} > 0$, and for any unit vector $\bar{\mathbf{s}} \in \mathcal{S}_k$, $J(\bar{\mathbf{s}}) \geq \bar{\epsilon}$. The desired result follows immediately. □

Having defined the securable and the unsecurable subspaces and established certain key properties of theirs, we now characterize a certain subspace of the watermark-securable subspace of system (8.3). Its operational meaning is established in Theorem 1 in the sequel.

Consider the set of all subspaces $\mathcal{C}$ such that

$$[\mathcal{R}(B^M) \oplus \mathcal{R}^\perp(B^H)]_{\mathcal{V}} \subseteq \mathcal{C} \subseteq \mathcal{V}, \tag{8.12}$$

and for every $\mathbf{c} \in \mathcal{C}$,

$$[A\mathbf{c}]_{\mathcal{C}^\perp \cap \mathcal{V}} = 0. \tag{8.13}$$

It is straightforward to verify that this set is closed under intersections. Define the subspace $\mathcal{G}_1$ as the smallest subspace satisfying (8.12) and (8.13), or more precisely, the intersection of all subspaces satisfying (8.12) and (8.13). Recall from Section 4.1 the notation $P_{\mathcal{X}}$, which denotes the projection matrix which projects a vector right-multiplying it onto the subspace $\mathcal{X}$, and define the subspace $\mathcal{G}_2$ as the span of the unstable subspace and the center subspace of the linear map $P_{\mathcal{G}_1^\perp \cap \mathcal{V}} A$.

Definition 8. *Define the subspace $\mathcal{W}$ of system (8.3) as*

$$\mathcal{W} \coloneqq (\mathcal{G}_1 \oplus \mathcal{G}_2)^\perp \cap \mathcal{V}. \tag{8.14}$$

As we show in Theorem 1, no matter what attack strategy the malicious nodes employ, they cannot degrade the state estimation performance of the honest nodes along the subspace $\mathcal{W}$, if they wish to remain undetected by Tests (8.4) and (8.5).

Lemma 14.

$$\mathcal{W} \perp \{\mathcal{R}(B^M) \oplus \mathcal{R}^{\perp}(B^H)\}. \tag{8.15}$$

Proof. We have from the definition of the watermark-securable subspace that $\mathcal{W} \perp \mathcal{G}_1$. We have from the definition of $\mathcal{G}_1$ that (8.12) holds with $\mathcal{C} = \mathcal{G}_1$, so that $\mathcal{G}_1 \supseteq [\mathcal{R}(B^M) \oplus \mathcal{R}^{\perp}(B^H)]_{\mathcal{V}}$. Combining these two relations, we have that $\mathcal{W} \perp [\mathcal{R}(B^M) \oplus \mathcal{R}^{\perp}(B^H)]_{\mathcal{V}}$. Let I_1 be a matrix whose columns form an orthonormal basis for the subspace $\mathcal{R}(B^M) \oplus \mathcal{R}^{\perp}(B^H)$ and I_2 be a matrix whose columns form an orthonormal basis for the subspace $\mathcal{W}$. Then, the previous relation implies that $I_2^T P_{\mathcal{V}} I_1 = 0$, which can be rewritten as $(P_{\mathcal{V}}^T I_2)^T I_1 = 0$. Since any orthogonal projection matrix is symmetric, we have that $P_{\mathcal{V}} = P_{\mathcal{V}}^T$. Using (8.14), we have that $\mathcal{W} \subseteq \mathcal{V}$, and hence, $(P_{\mathcal{V}}^T I_2) = I_2$. Consequently, $I_2^T I_1 = 0$, implying that $\mathcal{W} \perp \mathcal{R}(\mathcal{B}^{\mathcal{M}}) \oplus \mathcal{R}^{\perp}(\mathcal{B}^{\mathcal{H}})$. □

Let

$$\mathcal{H} \coloneqq \mathcal{G}_1 \oplus \mathcal{G}_2.$$

Lemma 15.

$$\mathcal{W} \perp A\mathcal{H}, \tag{8.16}$$

where $A\mathcal{H} \coloneqq \{A\mathbf{x} : \mathbf{x} \in \mathcal{H}\}$.

Proof. It follows from the definition of $\mathcal{G}_1$ that (8.13) holds with $\mathcal{C} = \mathcal{G}_1$, implying that

$$[A\mathcal{G}_1]_{\mathcal{G}_1^{\perp} \cap \mathcal{V}} = 0. \tag{8.17}$$

Using (8.14) and the identity that for any two subspaces $\mathcal{X}_1$ and $\mathcal{X}_2$, $[\mathcal{X}_1 \oplus \mathcal{X}_2]^{\perp} = \mathcal{X}_1^{\perp} \cap \mathcal{X}_2^{\perp}$, we

have $\mathcal{W} \subseteq \mathcal{G}_1^{\perp} \cap \mathcal{V}$. Combining this with (8.17) gives

$$\mathcal{W} \perp A\mathcal{G}_1. \tag{8.18}$$

Note next that the subspace $\mathcal{G}_2$ is $P_{\mathcal{G}_1^{\perp} \cap \mathcal{V}} A$-invariant, implying that

$$[A\mathcal{G}_2]_{\mathcal{G}_1^{\perp} \cap \mathcal{V}} \subseteq \mathcal{G}_2. \tag{8.19}$$

Now, $\mathcal{G}_1^{\perp} \cap \mathcal{V} = \mathcal{G}_2 \oplus \mathcal{W}$. To see this, note that since $\mathcal{G}_2$ is the span of the center and the unstable subspaces of $P_{\mathcal{G}_1^{\perp} \cap \mathcal{V}} A$, it follows that $\mathcal{G}_2 \subseteq \mathcal{G}_1^{\perp} \cap \mathcal{V}$. Let $\mathcal{B}_1$ be a matrix whose columns span the subspace $\mathcal{G}_1$, the matrix $\mathcal{B}_2$ be a matrix whose columns span the subspace $\mathcal{G}_2$, and $\mathcal{B}_{\mathcal{W}}$ be a matrix whose columns span the subspace $\mathcal{W}$. Then, it follows from (8.14) that columns of the matrix $[\mathcal{B}_1 \ \mathcal{B}_2 \ \mathcal{B}_{\mathcal{W}}]$ span the unsecurable subspace $\mathcal{V}$. The subspace $\mathcal{G}_1^{\perp} \cap \mathcal{V}$ is therefore spanned by the columns of the matrix $[\mathcal{B}_2 \ \mathcal{B}_{\mathcal{W}}]$, from which it follows immediately that $\mathcal{G}_1^{\perp} \cap \mathcal{V} = \mathcal{G}_2 \oplus \mathcal{W}$. Combining this with (8.19) yields $[A\mathcal{G}_2]_{\mathcal{W}} = 0$, i.e.,

$$\mathcal{W} \perp A\mathcal{G}_2. \tag{8.20}$$

Combining (8.18) and (8.20) implies (8.16). □

8.4 The case of Perfectly-Observed MIMO Systems

Consider system (8.3) and recall that $C = I$. Each honest sensor performs the tests (8.4) and (8.5). If the reported sequence of measurements $\{\mathbf{z}\}$ pass these tests, then there is no reason for the honest nodes to suspect that there are any malicious nodes in the system, and therefore, their estimate of the state at time k is simply $\mathbf{z}[k]$. The state estimation error $\mathbf{m}[t]$ incurred by the honest nodes at time t is therefore

$$\mathbf{m}[t] \coloneqq \mathbf{z}[t] - \mathbf{x}[t]. \tag{8.21}$$

The following theorem establishes that the projection of the state estimation error $\{\mathbf{m}\}$ along the subspace $\mathcal{S} \oplus \mathcal{W}$ can only be of zero power, no matter what attack strategy the malicious nodes employ.

Theorem 16. *Consider the system (8.3) and suppose that the reported measurements* $\{\mathbf{z}\}$ *pass tests (8.4) and (8.5) so that the malicious nodes remain undetected. Then,*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} \left|\left|\mathbf{m}_{\mathcal{S}\oplus\mathcal{W}}[k]\right|\right|^2 = 0. \tag{8.22}$$

Proof. We first show that the sequence $\{\mathbf{m}_{\mathcal{S}}\}$ is of zero power. Since for each honest sensor i, $i \in \{1, \ldots, p\}$, $z_i[k] \equiv x_i[k]$, we have that $m_i[k] \equiv 0$ for all such i. It follows that for a perfectly observed system,

$$\mathbf{m}_{\mathcal{S}_0}[k] \equiv 0. \tag{8.23}$$

Suppose now for induction that $\{\mathbf{m}_{\mathcal{S}_i}\}$ is of zero power for all $i \in \{0, \ldots, l\}$. I.e.,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} ||\mathbf{m}_{\mathcal{S}_i}[k]||^2 = 0 \tag{8.24}$$

for all $i \in \{0, \ldots, l\}$. We show that $\{\mathbf{m}_{\mathcal{S}_{l+1}}\}$ is also of zero power. We have from (8.4) that

$$\begin{aligned}\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (\mathbf{m}[k+1] - A\mathbf{m}[k] + B^H\mathbf{e}^H[k] + B^M\mathbf{e}^M[k] + \mathbf{w}[k+1])^{TT}_{[\oplus_{j=0}^{l}\mathcal{S}_j]\cap([\mathcal{R}(B^M)]_{\oplus_{j=0}^{l}\mathcal{S}_j})^{\perp}} \\ = P_l B B^T P_l^T \sigma_e^2 + \sigma_w^2 P_l P_l^T,\end{aligned} \tag{8.25}$$

where P_l is a projection matrix which projects a vector right-multiplying it to the subspace $[\oplus_{j=0}^{l}\mathcal{S}_j]\cap ([\mathcal{R}(B^M)]_{\oplus_{j=0}^{l}\mathcal{S}_j})^{\perp}$.

Now, write the vector $\mathbf{m}[t]$ as the sum of its components on the unsecurable subspace $\mathcal{V}$ and the subspaces $\mathcal{S}_0, \ldots, \mathcal{S}_{r-1}$, so that $\mathbf{m}[t] = \mathbf{m}_{\mathcal{V}}[t] + \sum_{j=0}^{r-1} \mathbf{m}_{\mathcal{S}_j}[t]$. It follows from Lemma 10 that

for each $i \in \{l+2, \ldots, r-1\}$, there exists $\mathbf{d}^i[k]$ such that

$$(-A\mathbf{m}_{\mathcal{S}_i}[k] + B^M \mathbf{d}^i[k])_{\oplus_{j=0}^{l}\mathcal{S}_j} = 0. \tag{8.26}$$

It also follows from the definition of the unsecurable subspace $\mathcal{V}$ that there exists $\mathbf{d}^{\mathcal{V}}[k]$ such that $-A\mathbf{m}_{\mathcal{V}}[k] + B^M \mathbf{d}^{\mathcal{V}}[k] \in \mathcal{V}$, and therefore,

$$(-A\mathbf{m}_{\mathcal{V}}[k] + B^M \mathbf{d}^{\mathcal{V}}[k])_{\oplus_{j=0}^{l}\mathcal{S}_j} = 0. \tag{8.27}$$

Define $\mathbf{d}^U[k] \coloneqq \mathbf{e}^M[k] - \mathbf{d}^{\mathcal{V}}[k] - \sum_{i=l+2}^{r-1} \mathbf{d}^i[k]$. Substituting this in (8.25) gives

$$\begin{aligned}
\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} \Big(\sum_{i=0}^{r-1} \mathbf{m}_{\mathcal{S}_i}[k+1] + \mathbf{m}_{\mathcal{V}}[k+1] - \sum_{i=0}^{l} A\mathbf{m}_{\mathcal{S}_i}[k] \\
- A\mathbf{m}_{\mathcal{S}_{l+1}}[k] + B^M \mathbf{d}^U[k] \\
+ \sum_{i=l+2}^{r-1} (-A\mathbf{m}_{\mathcal{S}_i}[k] + B^M \mathbf{d}^i[k]) - A\mathbf{m}_{\mathcal{V}}[k] + B^M \mathbf{d}^{\mathcal{V}}[k] \\
+ B^H \mathbf{e}^H[k] + \mathbf{w}[k+1] \Big)^{TT}_{[\oplus_{j=0}^{l}\mathcal{S}_j] \cap ([\mathcal{R}(B^M)]_{\oplus_{j=0}^{l}\mathcal{S}_j})^{\perp}} \\
= P_l B B^T P_l^T \sigma_e^2 + \sigma_w^2 P_l P_l^T.
\end{aligned} \tag{8.28}$$

Using (8.26) and (8.27), the above simplifies to

$$\begin{aligned}
\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} \Big(\sum_{i=0}^{l} \mathbf{m}_{\mathcal{S}_i}[k+1] - \sum_{i=0}^{l} A\mathbf{m}_{\mathcal{S}_i}[k] \\
- A\mathbf{m}_{\mathcal{S}_{l+1}}[k] + B^M \mathbf{d}^U[k] \\
+ B^H \mathbf{e}^H[k] + \mathbf{w}[k+1] \Big)^{TT}_{[\oplus_{j=0}^{l}\mathcal{S}_j] \cap ([\mathcal{R}(B^M)]_{\oplus_{j=0}^{l}\mathcal{S}_j})^{\perp}} \\
= P_l B B^T P_l^T \sigma_e^2 + \sigma_w^2 P_l P_l^T.
\end{aligned} \tag{8.29}$$

The induction hypothesis implies that the sequence $\{\sum_{i=0}^{l} \mathbf{m}_{\mathcal{S}_i}[k+1] - \sum_{i=0}^{l} A\mathbf{m}_{\mathcal{S}_i}[k]\}$ is of zero

power. Therefore, the above equality, after some algebra, further simplifies to

$$\begin{aligned}
\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}&(-A\mathbf{m}_{\mathcal{S}_{l+1}}[k]+B^M\mathbf{d}^U[k])^{TT}_{[\oplus_{j=0}^{l}\mathcal{S}_j]\cap([\mathcal{R}(B^M)]_{\oplus_{j=0}^{l}\mathcal{S}_j})^\perp}\\
&+(-A\mathbf{m}_{\mathcal{S}_{l+1}}[k]+B^M\mathbf{d}^U[k])_{[\oplus_{j=0}^{l}\mathcal{S}_j]\cap([\mathcal{R}(B^M)]_{\oplus_{j=0}^{l}\mathcal{S}_j})^\perp}\\
&\quad\times(B^H\mathbf{e}^H[k]+\mathbf{w}[k+1])^{T}_{[\oplus_{j=0}^{l}\mathcal{S}_j]\cap([\mathcal{R}(B^M)]_{\oplus_{j=0}^{l}\mathcal{S}_j})^\perp}\\
&+(B^H\mathbf{e}^H[k]+\mathbf{w}[k+1])_{[\oplus_{j=0}^{l}\mathcal{S}_j]\cap([\mathcal{R}(B^M)]_{\oplus_{j=0}^{l}\mathcal{S}_j})^\perp}\\
&\quad\times(-A\mathbf{m}_{\mathcal{S}_{l+1}}[k]+B^M\mathbf{d}^U[k])^{T}_{[\oplus_{j=0}^{l}\mathcal{S}_j]\cap([\mathcal{R}(B^M)]_{\oplus_{j=0}^{l}\mathcal{S}_j})^\perp}\\
&\quad=P_lB^MB^{M^T}P_l^T\sigma_e^2. \qquad (8.30)
\end{aligned}$$

Since $[\mathcal{R}(B^M)]_{[\oplus_{j=0}^{l}\mathcal{S}_j]\cap([\mathcal{R}(B^M)]_{\oplus_{j=0}^{l}\mathcal{S}_j})^\perp}=\{0\}$, we have that $P_lB^M=0$, and consequently, the Right Hand Side (RHS) of the above equality reduces to zero. Similarly, the vector $B^M\mathbf{d}^U[k]$ projected on the subspace $([\mathcal{R}(B^M)]_{\oplus_{j=0}^{l}\mathcal{S}_j})^\perp$ equals zero, and consequently, the Left Hand Side (LHS) also simplifies, and we have

$$\begin{aligned}
\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}&(-A\mathbf{m}_{\mathcal{S}_{l+1}}[k])^{TT}_{[\oplus_{j=0}^{l}\mathcal{S}_j]\cap([\mathcal{R}(B^M)]_{\oplus_{j=0}^{l}\mathcal{S}_j})^\perp}\\
&+(-A\mathbf{m}_{\mathcal{S}_{l+1}}[k])_{[\oplus_{j=0}^{l}\mathcal{S}_j]\cap([\mathcal{R}(B^M)]_{\oplus_{j=0}^{l}\mathcal{S}_j})^\perp}(B^H\mathbf{e}^H[k]+\mathbf{w}[k+1])^{T}_{[\oplus_{j=0}^{l}\mathcal{S}_j]\cap([\mathcal{R}(B^M)]_{\oplus_{j=0}^{l}\mathcal{S}_j})^\perp}\\
&+(B^H\mathbf{e}^H[k]+\mathbf{w}[k+1])_{[\oplus_{j=0}^{l}\mathcal{S}_j]\cap([\mathcal{R}(B^M)]_{\oplus_{j=0}^{l}\mathcal{S}_j})^\perp}(-A\mathbf{m}_{\mathcal{S}_{l+1}}[k])^{T}_{[\oplus_{j=0}^{l}\mathcal{S}_j]\cap([\mathcal{R}(B^M)]_{\oplus_{j=0}^{l}\mathcal{S}_j})^\perp}\\
&=0. \qquad (8.31)
\end{aligned}$$

For brevity of notation, let

$$(-A\mathbf{m}_{\mathcal{S}_{l+1}}[k])_{[\oplus_{j=0}^{l}\mathcal{S}_j]\cap([\mathcal{R}(B^M)]_{\oplus_{j=0}^{l}\mathcal{S}_j})^\perp}=:\boldsymbol{\alpha}[k],$$

and

$$(B^H\mathbf{e}^H[k]+\mathbf{w}[k+1])_{[\oplus_{j=0}^{l}\mathcal{S}_j]\cap([\mathcal{R}(B^M)]_{\oplus_{j=0}^{l}\mathcal{S}_j})^\perp} \coloneqq \boldsymbol{\beta}[k+1],$$

so that (8.31) becomes

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}\boldsymbol{\alpha}^{TT}[k]+\boldsymbol{\alpha}[k]\boldsymbol{\beta}^T[k+1]+\boldsymbol{\beta}[k+1]\boldsymbol{\alpha}^T[k]=0. \tag{8.32}$$

Define the $\sigma-$algebra $\mathcal{F}_k \coloneqq \sigma(\mathbf{z}^k,\mathbf{x}^k,\mathbf{e}^{M\,k})$. Then, $-A\mathbf{m}_{\mathcal{S}_{l+1}}[k]\in\mathcal{F}_k$, and therefore $\boldsymbol{\alpha}[k]\in\mathcal{F}_k$. Also, $(B^H\mathbf{e}^H[k]+\mathbf{w}[k+1],\mathcal{F}_{k+1})$ is a Martingale Difference Sequence, and so is $(\boldsymbol{\beta}[k+1],\mathcal{F}_{k+1})$. The Martingale Stability Theorem (MST) [49] applies, and equating the i^{th} entry along the diagonal of (8.32) yields

$$\sum_{k=0}^{T-1}\alpha_i^2[k]+o(\sum_{k=0}^{T-1}\alpha_i^2[k])+O(1)=o(T). \tag{8.33}$$

Dividing the above equality by T and taking the limit as $T\to\infty$ implies that $\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}\alpha_i^2[k]=0$ for all $i\in\{1,\ldots,p\}$, which in turn implies that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}||(-A\mathbf{m}_{\mathcal{S}_{l+1}}[k])_{[\oplus_{j=0}^{l}\mathcal{S}_j]\cap([\mathcal{R}(B^M)]_{\oplus_{j=0}^{l}\mathcal{S}_j})^\perp}||^2=0. \tag{8.34}$$

Combining the above equality with (8.11) gives

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}\left|\left|\mathbf{m}_{\mathcal{S}_{l+1}}[k]\right|\right|^2=0, \tag{8.35}$$

thereby completing the induction.

We next show that the sequence $\{\mathbf{m}_{\mathcal{W}}\}$ is of zero power. Define

$$\mathbf{v}[t+1]\coloneqq\mathbf{z}[t+1]-A\mathbf{z}[t]-Bg_t(\mathbf{z}^t)-B^H\mathbf{e}^H[t]-B^M\mathbf{e}^M[t]-\mathbf{w}[t+1], \tag{8.36}$$

It follows from (8.5) that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}\mathbf{e}^H[k](B^H\mathbf{e}^H[k]+B^M\mathbf{e}^M[k]+\mathbf{w}[k+1]+\mathbf{v}[k+1])^T=\sigma_e^2B^{H^T},$$

whence

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}\mathbf{e}^H[k](B^H\mathbf{e}^H[k]+B^M\mathbf{e}^M[k]+\mathbf{w}[k+1]+\mathbf{v}[k+1])^T_{\mathcal{W}}=\sigma_e^2B^{H^T}P^T_{\mathcal{W}},$$

where $P_{\mathcal{W}}$ denotes the matrix which projects a vector right-multiplying it onto the subspace $\mathcal{W}$. Using (8.15), the left-hand side (LHS) of the above equality simplifies, and we have

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}\mathbf{e}^H[k](P_{\mathcal{W}}B^H\mathbf{e}^H[k]+\mathbf{w}_{\mathcal{W}}[k+1]+\mathbf{v}_{\mathcal{W}}[k+1])^T=\sigma_e^2B^{H^T}P^T_{\mathcal{W}},$$

which in turn yields

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}\mathbf{e}^H[k]\mathbf{v}^T_{\mathcal{W}}[k+1]=0. \tag{8.37}$$

It follows from (8.4) that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}(B^H\mathbf{e}^H[k]+B^M\mathbf{e}^M[k]\mathbf{w}[k+1]+\mathbf{v}[k+1])^{TT}_{\mathcal{W}}=P_{\mathcal{W}}BB^TP^T_{\mathcal{W}}\sigma_e^2+\sigma_w^2P_{\mathcal{W}}P^T_{\mathcal{W}},$$

which using (8.15) simplifies to

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}(P_{\mathcal{W}}B^H\mathbf{e}^H[k]+\mathbf{w}_{\mathcal{W}}[k+1]+\mathbf{v}_{\mathcal{W}}[k+1])^{TT}=P_{\mathcal{W}}B^HB^{H^T}P^T_{\mathcal{W}}\sigma_e^2+\sigma_w^2P_{\mathcal{W}}P^T_{\mathcal{W}}.$$

Using (8.37), the above equality further simplifies to

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}\mathbf{w}_{\mathcal{W}}[k+1]\mathbf{v}^T_{\mathcal{W}}[k+1]+\mathbf{v}_{\mathcal{W}}\quad[k+1]\mathbf{w}^T_{\mathcal{W}}[k+1]+\mathbf{v}_{\mathcal{W}}[k+1]\mathbf{v}^T_{\mathcal{W}}[k+1]=0. \tag{8.38}$$

The malicious sensors can adapt the measurements that they report at any time t to all past observations as well as all past reported measurements. In addition, we also allow the malicious sensors to employ an additional randomization in reporting their measurements. Let $\boldsymbol{\theta}[t]$ be a random variable independent of all measurements observed up to time t and all measurements reported up to time $t-1$. We allow the measurement $\mathbf{z}[t]$ that is reported at time t to also depend on $\boldsymbol{\theta}[t]$. Define $\mathcal{F}_k \coloneqq \sigma(\mathbf{x}^k, \boldsymbol{\theta}^k, \mathbf{e}^{M^{k-1}}, \mathbf{e}^{H^{k-2}})$, $\widehat{\mathbf{w}}[k+1] \coloneqq \mathbb{E}[\mathbf{w}[k+1]|\mathcal{F}_{k+1}]$, $\widehat{\mathbf{w}}_{\mathcal{R}(B^H)}[k+1] \coloneqq \mathbb{E}[\mathbf{w}_{\mathcal{R}(B^H)}[k+1]|\mathcal{F}_{k+1}]$. The quantities $\widehat{\mathbf{w}}_{\mathcal{R}^\perp(B^H)}[k+1]$ and $\widehat{\mathbf{w}}_{\mathcal{W}}[k+1]$ are defined likewise. Now,

$$\mathbf{w}_{\mathcal{W}}[k+1] = P_{\mathcal{W}}\mathbf{w}[k+1] = P_{\mathcal{W}}\mathbf{w}_{\mathcal{R}(B^H)}[k+1] + P_{\mathcal{W}}\mathbf{w}_{\mathcal{R}^\perp(B^H)}[k+1] = P_{\mathcal{W}}\mathbf{w}_{\mathcal{R}(B^H)}[k+1], \tag{8.39}$$

where the last equality follows from (8.15). Let $B^H = Q^H R^H$ be the QR-decomposition of the matrix B^H. I.e., the columns of the matrix Q^H form an orthonormal basis for the subspace $\mathcal{R}(B^H)$ and the matrix R^H is upper-triangular. We then have that for every $k+1$, there exists a vector $\mathbf{w}^c[k+1]$ such that

$$\mathbf{w}_{\mathcal{R}(B^H)}[k+1] = Q^H \mathbf{w}^c[k+1]. \tag{8.40}$$

We have from (8.3) that $\mathbf{x}[k+1] - A\mathbf{x}[k] - Bg_k(\mathbf{z}^k) - B^M\mathbf{e}^M[k] = B^H\mathbf{e}^H[k] + \mathbf{w}[k+1] = Q^H R^H \mathbf{e}^H[k] + Q^H\mathbf{w}^c[k+1] + \mathbf{w}_{\mathcal{R}^\perp(B^H)}[k+1]$, implying that $Q^{H^T}(\mathbf{x}[k+1] - A\mathbf{x}[k] - Bg_k(\mathbf{z}^k) - B^M\mathbf{e}^M[k]) = R^H\mathbf{e}^H[k] + \mathbf{w}^c[k+1]$. It follows that $\mathbb{E}[\mathbf{w}^c[k+1]|\mathcal{F}_{k+1}] = \sigma_w^2(\sigma_e^2 R^H R^{H^T} + \sigma_w^2 I)^{-1}(Q^{H^T}(\mathbf{x}[k+1] - A\mathbf{x}[k] - Bg_k(\mathbf{z}^k) - B^M\mathbf{e}^M[k]))$, whence

$$\widehat{\mathbf{w}}^c[k+1] \coloneqq \mathbb{E}[\mathbf{w}^c[k+1]|\mathcal{F}_{k+1}] = K_W(R^H\mathbf{e}^H[k] + \mathbf{w}^c[k+1]), \tag{8.41}$$

where $K_W \coloneqq \sigma_w^2(\sigma_e^2 R^H R^{H^T} + \sigma_w^2 I)^{-1}$ has all eigenvalues strictly lesser than unity. Define $\widetilde{\mathbf{w}}^c[k+1] \coloneqq \mathbf{w}^c[k+1] - \widehat{\mathbf{w}}[k+1]$, so that $\mathbf{w}^c[k+1] = \widehat{\mathbf{w}}^c[k+1] + \widetilde{\mathbf{w}}^c[k+1]$. Substituting (8.41) in the

above equality, we have $\mathbf{w}^c[k+1] = K_W R^H \mathbf{e}^H[k] + K_W \mathbf{w}^c[k+1] + \widetilde{\mathbf{w}}^c[k+1]$, whence

$$\mathbf{w}^c[k+1] = (I - K_W)^{-1} K_W R^H \mathbf{e}^H[k] + (I - K_W)^{-1} \widetilde{\mathbf{w}}^c[k+1].$$

Substituting this in (8.40) gives

$$\mathbf{w}_{\mathcal{R}(B^H)}[k+1] = Q^H (I - K_W)^{-1} K_W R^H \mathbf{e}^H[k] + Q^H (I - K_W)^{-1} \widetilde{\mathbf{w}}^c[k+1],$$

and substituting this in (8.39) gives

$$\mathbf{w}_{\mathcal{W}}[k+1] = P_{\mathcal{W}} Q^H (I - K_W)^{-1} K_W R^H \mathbf{e}^H[k] + P_{\mathcal{W}} Q^H (I - K_W)^{-1} \widetilde{\mathbf{w}}^c[k+1]. \tag{8.42}$$

Now, $(\widetilde{\mathbf{w}}^c[k], \mathcal{F}_{k+1})$ is a martingale difference sequence, and $\mathbf{v}[k+1] \in \mathcal{F}_{k+1}$. Using the Martingale Stability Theorem [49], we have

$$\sum_{k=0}^{T-1} \widetilde{\mathbf{w}}^c[k+1] \mathbf{v}_{\mathcal{W}}^T[k+1] = \begin{bmatrix} o(\sum_{k=0}^{T-1} v_{\mathcal{W}1}^2[k+1]) & \dots & o(\sum_{k=0}^{T-1} v_{\mathcal{W}p}^2[k+1]) \\ \vdots & \ddots & \vdots \\ o(\sum_{k=0}^{T-1} v_{\mathcal{W}1}^2[k+1]) & \dots & o(\sum_{k=0}^{T-1} v_{\mathcal{W}p}^2[k+1]) \end{bmatrix} \tag{8.43}$$

Substituting (8.42) in (8.38), then using (8.37) and (8.43), and then equating the trace yields

$$\lim_{T \to \infty} \frac{1}{T} \sum_{k=0}^{T-1} \left|\left| \mathbf{v}_{\mathcal{W}}[k+1] \right|\right|^2 = 0. \tag{8.44}$$

We have from (8.36) and (8.3) that

$$\mathbf{v}[t+1] = \mathbf{m}[t+1] - A\mathbf{m}[t] = \mathbf{m}_{\mathcal{S}}[t+1] + \mathbf{m}_{\mathcal{H}}[t+1] + \mathbf{m}_{\mathcal{W}}[t+1] \\ - A\mathbf{m}_{\mathcal{S}}[t] - A\mathbf{m}_{\mathcal{H}}[t] - A\mathbf{m}_{\mathcal{W}}[t].$$

Projecting $\mathbf{v}[t+1]$ onto the subspace $\mathcal{W}$ and using (8.17) gives $\mathbf{v}_{\mathcal{W}}[t+1] = \mathbf{m}_{\mathcal{W}}[t+1] -$

$P_{\mathcal{W}}A\mathbf{m}_{\mathcal{S}}[t] - P_{\mathcal{W}}A\mathbf{m}_{\mathcal{W}}[t]$, which upon rearranging yields

$$\mathbf{m}_{\mathcal{W}}[t+1] = P_{\mathcal{W}}A\mathbf{m}_{\mathcal{W}}[t] + P_{\mathcal{W}}A\mathbf{m}_{\mathcal{S}}[t] + \mathbf{v}_{\mathcal{W}}[t+1].$$

Using (8.44), the fact that $\{\mathbf{m}_{\mathcal{S}}\}$ is of zero power, and the stability of the matrix $P_{\mathcal{W}}A$, we have that $\{\mathbf{m}_{\mathcal{W}}\}$ is of zero power. □

8.5 Conclusion

In prior chapters, the approach of Dynamic Watermarking was established as providing security guarantees against arbitrary sensor attacks. In this chapter, we have considered a system which includes malicious sensors and actuators mixed with honest sensors and actuators, and wherein the honest actuators employ Dynamic Watermarking. Given any such system, we have characterized a certain subspace of the watermark-securable subspace, and have established that the malicious sensors and actuators cannot degrade the state estimation performance of the honest sensors and actuators along this subspace.

9. THEORY AND IMPLEMENTATION OF DYNAMIC WATERMARKING FOR CYBERSECURITY OF ADVANCED TRANSPORTATION SYSTEMS*

9.1 Introduction

Recently there has been great interest in automated and semi-automated transportation systems involving various driver assists. These advanced systems rely on sensors to provide state information and situational awareness to the control logic governing the vehicle. However, this has also increased the vulnerability of these advanced transportation systems to cyber attacks. In fact, there have been demonstrated cyber attacks on automobiles in the recent past [64, 65], where two hackers have remotely subverted an automobile, taking control of its steering and braking units. This ultimately led to the automobile manufacturer recalling over a million cars to patch the identified vulnerabilities. Several other similar news reports [66, 67, 68] point the need for cybersecurity of automated transportation systems. In this chapter we demonstrate how the technique of dynamic watermarking can be employed to secure an automated transportation system against arbitrary attacks on its sensors. This chapter is based on joint work with Dr. Woo-Hyun Ko of Texas A&M University.

In the previous chapter, the technique of dynamic watermarking of signals has been proposed to secure such cyberphysical systems. It has been shown that in theory, employing dynamic watermarking and using two specific tests can detect erroneous sensor measurements for a large class of linear systems. In this chapter, we investigate whether this method can actually be used in a real transportation system to detect attacks on the positioning sensor systems and thereby prevent collisions. For this purpose we extend the theory to allow for several nonlinearities not accounted for in an idealized linear system on which the theoretical results have previously been established. After

*Reprinted with permission from “Theory and implementation of dynamic watermarking for cybersecurity of advanced transportation systems," by W-H. Ko, B. Satchidanandan and P. R. Kumar in *Proceedings of the 2016 IEEE Conference on Communications and Network Security (CNS),* pp. 416-420, 2016, IEEE.

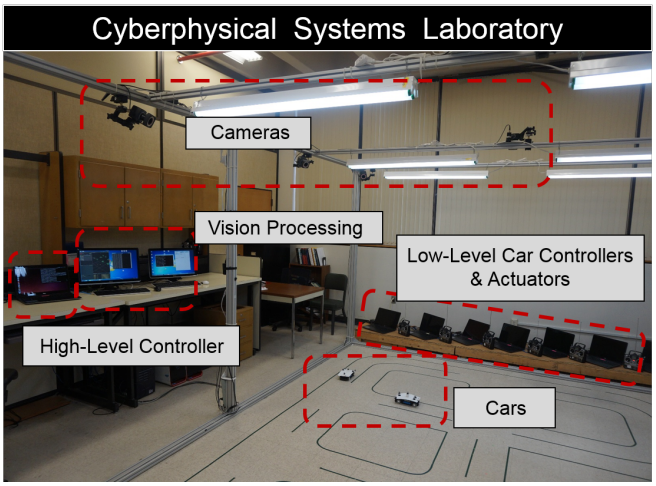

Figure 9.1: Experimental Facility for Dynamic Watermarking. Reprinted with permission from [2].

doing so, we have implemented the dynamic watermarking method on a laboratory autonomous transportation system. We demonstrate in this chapter that this method successfully detects and responds to attacks on the position sensors which can potentially cause collisions, thereby preventing collisions.

9.2 Attacking an autonomous transportation system by malicious attacks on sensors

Advanced transportation systems, whether fully autonomous or those that provide driver assist, employ sensors to determine the position of a vehicle. Such sensors can be maliciously attacked. We first begin by showing how such an attack can be used to create collisions. The experimental setup, shown in Fig. 9.1, is a prototype of an intelligent transportation system housed in the Cyberphysical Systems Laboratory of Texas A&M University. At the core of the system is a set of vehicles that are required to follow a particular trajectory within a rectangular area. This can be thought of as the high-level control objective. The system consists of a supervisory layer or a high-level controller which decides the trajectory that each vehicle should follow.

Attached to the setup is a set of ten cameras which capture an image of the rectangular area along with the vehicles once every 100ms. These images are transmitted to the vision sensors in the system which accurately compute, from these raw images, the coordinates (x_i, y_i) and orientation

θ_i of vehicle i at each sampling instant t. The vision sensors then transmit this information to the vision server, which disseminates this information to other modules in the system such as the low-level controller, supervisor, collision avoidance module, etc.

The low-level controller determines the control input to be applied to each vehicle using Model Predictive Control. It computes the control input using the position and orientation information of each vehicle that it obtains from the vision server, and also their reference trajectories that it obtains from the supervisor. The controller then sends this information to the collision avoidance module.

The collision avoidance module uses this control input, the position estimates provided by the vision server, and the dynamic model of the vehicles to predict their position during the next sampling epoch. If it detects a collision based on this computation, it instructs the actuator to halt the vehicles. If not, it relays the control input computed by the controller as such to the actuator, which then applies that particular input.

The plant model for vehicle i is given by its kinematic equations:

$$x_i[t+1] = x_i[t] + h\cos(\theta_i[t])v_i[t] + h\cos(\theta_i[t])w_{ix}[t], \tag{9.1}$$

$$y_i[t+1] = y_i[t] + h\sin(\theta_i[t])v_i[t] + h\sin(\theta_i[t])w_{iy}[t], \tag{9.2}$$

$$\theta_i[t+1] = \theta_i[t] + h\omega_i[t] + hw_{i\theta}[t], \tag{9.3}$$

where h is the sampling time period (100ms in this case), $v_i[t]$ is the speed of the vehicle at sampling epoch t and is one of the control inputs of the vehicles, and $\omega_i[t]$ is the angular speed of the vehicle at sampling epoch t and is the second control input of the vehicles. Also, $w_{ix}[t]$, $w_{iy}[t]$, and $w_{i\theta}[t]$ are random variables whose variances we denote by σ_x^2, σ_y^2, and σ_θ^2 respectively, and they model the ambient noise entering the system as a consequence of small, random drifts in the actual values of the applied control inputs from their set points. We model them as zero-mean, i.i.d. normal random variables. For the purposes of our demonstration, it suffices to use just two vehicles, so that $i \in \{1, 2\}$. We make both of them follow an elliptical trajectory, one behind the

other. The attached video clip opens by showing the vehicle trajectories in the absence of both attacks and dynamic watermarking.

Our system includes a collision avoidance module [69] that halts the vehicles when an imminent collision is detected. To illustrate its behavior, when a manually controlled vehicle is made to intercept the two vehicles' trajectories, the collision avoidance module detects imminent collisions and commands the actuators to halt the vehicles, thereby avoiding collisions. In a prior work [69], it was shown that this system indeed guarantees collision freedom.

Next, we construct a specific attack which spoofs the collision avoidance module by sending malicious position information to it. Specifically, we introduce maliciousness in the vision sensor which computes the $x-$coordinate of the vehicles' position from the image that it receives from the cameras. The attack strategy then is as follows. Let t_A denote the time at which the attack begins. Then, $z_{2x}[t_A] = x_2[t_A] + \tau$, where τ is the bias that the sensor adds to the $x-$coordinate of the vehicle. For $t > t_A$, the malicious sensor reports measurements $\{z_{2x}\}$ generated as

$$z_{2x}[t+1] = z_{2x}[t] + h\cos(\theta_2[t])u_2^g(\mathbf{z}_1^t, \mathbf{z}_2^t) + \cos(\theta_2[t])n[t], \tag{9.4}$$

where $n[t] \sim \mathcal{N}(0, \sigma_x^2)$. Therefore, once the attack begins, wrong position information is sent to the vision server, and consequently to all other modules in the system. In particular, wrong position information is sent to the collision avoidance module, which results in it not detecting imminent collisions. The attached clip demonstrates this attack which culminates in the two vehicles colliding with each other.

9.3 Protecting transportation system from malicious attacks on sensors through a nonlinear extension of dynamic watermarking

We now consider how the transportation system can be protected malicious attacks on sensors. We first extend the theory of dynamic watermarking to nonlinear systems to address the equations of vehicular motion. Consider a controller, as in Section-9.2, which computes the control policy-

specified input and superimposes the watermark on it. The system evolves as

$$x_i[t+1] = x_i[t] + h\cos(\theta_i[t])u_i^g(\mathbf{z}_1^t, \mathbf{z}_2^t) + h\cos(\theta_i[t])e_{iv}[t] + h\cos(\theta_i[t])w_{ix}[t], \tag{9.5}$$

$$y_i[t+1] = y_i[t] + h\sin(\theta_i[t])u_i^g(\mathbf{z}_1^t, \mathbf{z}_2^t) + h\sin(\theta_i[t])e_{iv}[t] + h\sin(\theta_i[t])w_{iy}[t], \tag{9.6}$$

$$\theta_i[t+1] = \theta_i[t] + h\omega_i[t] + he_{i\theta}[t] + hw_{i\theta}[t], \tag{9.7}$$

where $e_{iv}[t] \sim \mathcal{N}(0, \sigma_e^2)$ and i.i.d. across time is the watermark superimposed on the translation velocity control input $v_i[t]$, $e_{i\theta}[t] \sim \mathcal{N}(0, \sigma_\theta^2)$ and i.i.d. across time is the watermark superimposed on the angular velocity control input $\omega_i[t]$, $\mathbf{z}_i[t] = [z_{ix}[t], z_{ix}[t], z_{ix}[t]] = [z_{ix}[t], y_i[t], \theta_i[t]]^T$, and $z_{ix}[t]$ is vehicle-i's $x-$coordinate reported by the vision sensor at time t, which is different from its true value $x_i[t]$, $z_{iy}[k]$ and $z_{i\theta}[k]$ are the values reported for $y_i[k]$ and $\theta_i[k]$ respectively, which are equal to their true values at all times. The controller does not know a priori which of the sensors are malicious, if any.

The nominal trajectory of the vehicles in the presence and absence of dynamic watermarking are compared in Fig. 9.2. The mismatch between the nominal and actual trajectory is negligible, and in principle, can be made arbitrarily small.

The controller performs the following tests to check for maliciousness. The tests are specified only for $\{z_{ix}\}$ below, but analogous tests are also carried out by the controller for $\{z_{iy}\}$ and $\{z_{i\theta}\}$. While we state the following as standalone tests, they are subsumed by (6.70) and (6.71).

1. **Test 1:** The controller checks if

$$\lim_{t\to\infty} \frac{1}{t}\sum_{k=0}^{t-1}(z_{ix}[k+1] - z_{ix}[k] - h\cos(z_{i\theta}[k])u_i^g(\mathbf{z}_1^t, \mathbf{z}_2^t) - h\cos(z_{i\theta}[k])e_{iv}[k])^2 = \widetilde{\sigma}_x^2. \tag{9.8}$$

2. **Test 2:** The controller checks if

$$\lim_{t\to\infty} \frac{1}{t}\sum_{k=0}^{t-1}(z_{ix}[k+1] - z_{ix}[k] - h\cos(z_{i\theta}[k])u_i^g(\mathbf{z}_1^t, \mathbf{z}_2^t))^2 = \sigma_c^2, \tag{9.9}$$

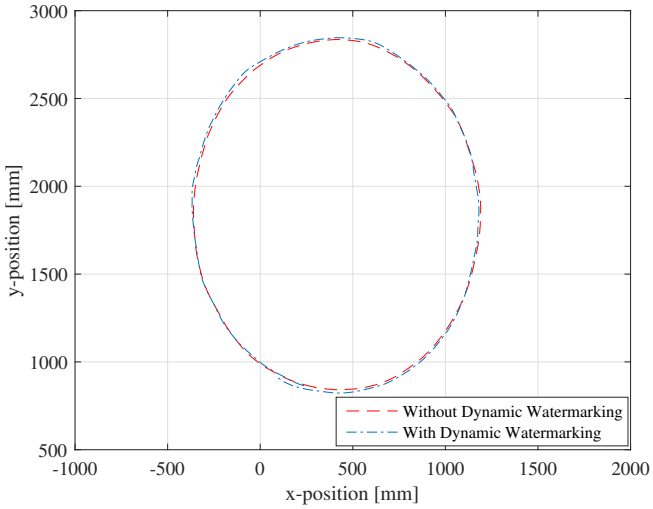

Figure 9.2: Trajectories of the vehicles with and without Dynamic Watermarking. Reprinted with permission from [2].

where

$$\widetilde{\sigma}_x^2 := \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} (h\cos(\theta_i[k])w_{ix}[k])^2, \tag{9.10}$$

and

$$\sigma_c^2 := \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} (h\cos(\theta_i[k])e_{iv}[k] + h\cos(\theta_i[k])w_{ix}[k])^2 \tag{9.11}$$

are the values that would be obtained for the LHS of (9.8) and (9.9) respectively had $\{z_{ix}\}$ been equal to $\{x_i\}$. The above quantities are the sum of independent, but non-identically distributed random variables. We assume that for the trajectory followed by the vehicles, the above limits exist. Fig. 9.3 and Fig. 9.4) plot $\frac{1}{t}\sum_{k=0}^{t-1}(h\cos(\theta_i[k])w_{ix}[k])^2$ and $\frac{1}{t}\sum_{k=0}^{t-1}(h\cos(\theta_i[k])e_{iv}[k] + h\cos(\theta_i[k])w_{ix}[k])^2$ as a function of time, and support our assumption that the limits (9.10) and (9.11) exist. These quantities were computed experimentally as follows in our demonstration.

Since in the absence of attacks, we have

$$\widetilde{\sigma}_x^2 = \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} x_i[k+1] - x_i[k] - h\cos(\theta_i[k]) u_i^g(\mathbf{z}_1^k, \mathbf{z}_2^k) - h\cos(\theta_i[k]) e_{iv}[k],$$

and

$$\sigma_c^2 = \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} x_i[k+1] - x_i[k] - h\cos(\theta_i[k]) u_i^g(\mathbf{z}_1^k, \mathbf{z}_2^k),$$

evaluating the RHS of the above equations from the experiment in the absence of attacks yield the desired noise variances.

The following theorem ensures that the the above two tests are sufficient to restrict the malicious sensor to adding an additive distortion that can only have zero power asymptotically.

Theorem 17. *Define*

$$v_x[t+1] \coloneqq z_{2x}[t+1] - z_{2x}[t] - h\cos(\theta_2[t]) u_2^g(\mathbf{z}_1^t, \mathbf{z}_2^t) - h\cos(\theta_2[t]) e_{2v}[t] - h\cos(\theta_2[t]) w_{2x}[t], \tag{9.12}$$

so that for an honest sensor, $v_x \equiv 0$. *If the reported sequence of measurements satisfy (9.8) and (9.9), then,*

$$\lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} v_x^2[k+1] = 0 \tag{9.13}$$

Proof. Since the sequence of reported measurements $\{z_{2x}\}$ satisfy (9.8), we have using (9.12)

$$\lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t} (h\cos(\theta_2[k]) w_{2x}[k] + v_x[k+1])^2 = \widetilde{\sigma}_x^2. \tag{9.14}$$

Expanding the above and using (9.10), we get

$$\lim_{t\to\infty}\frac{1}{t}\sum_{k=0}^{t} v_x^2[k+1] + 2h\cos(\theta_2[k])w_{2x}[k]v_x[k+1] = 0. \tag{9.15}$$

Similarly, since the reported measurements satisfy (9.9) we have,

$$\lim_{t\to\infty}\frac{1}{t}\sum_{k=0}^{t}(h\cos(\theta_2[k])w_{2x}[k] + h\cos(\theta_2[k])e_{2v}[k] + v_x[k+1])^2 = \sigma_c^2. \tag{9.16}$$

Expanding the above and using (9.10), (9.11), and (9.15), we arrive at

$$\lim_{t\to\infty}\frac{1}{t}\sum_{k=0}^{t}\cos(\theta_2[k])e_{2v}[k]v_x[k+1] = 0. \tag{9.17}$$

Now, define $\mathcal{F}_k \coloneqq \sigma(x^k, y^{k-1}, \theta^{k-1}, e_{2v}^{k-2}, z^k)$, $\widehat{w}[k] \coloneqq E[w[k]\big|\mathcal{F}_{k+1}]$, and $\widetilde{w}[k] \coloneqq w[k] - \widehat{w}[k]$. Then, for k such that $\cos(\theta_2[k]) \neq 0$, we have $\widehat{w}[k] = \frac{\sigma_w^2}{\sigma_e^2+\sigma_w^2}(e[k] + w[k])$. Now,

$$\sum_{k=0}^{t-1}\cos(\theta_2[k])w_{2x}[k]v_x[k+1] = \sum_{\substack{k=0,\\ \cos(\theta_2[k])\neq 0}}^{t-1}\cos(\theta_2[k])w_{2x}[k]v_x[k+1].$$

Expressing $w_{2x}[k]$ as $\widehat{w}[k] + \widetilde{w}[k]$, substituting for $\widehat{w}[k]$ using the aforementioned expression, and rearranging the terms gives

$$\begin{aligned}\sum_{\substack{k=0,\\ \cos(\theta_2[k])\neq 0}}^{t-1}\cos(\theta_2[k])w_{2x}[k]v_x[k+1] = \frac{\beta}{1-\beta}\sum_{\substack{k=0,\\ \cos(\theta_2[k])\neq 0}}^{t-1}\cos(\theta_2[k])e_{2v}[k]v_x[k+1] \\ + \frac{1}{1-\beta}\sum_{\substack{k=0,\\ \cos(\theta_2[k])\neq 0}}^{t-1}\cos(\theta_2[k])\widetilde{w}_{2x}[k]v_x[k+1],\end{aligned} \tag{9.18}$$

where $\beta \coloneqq \frac{\sigma_w^2}{\sigma_e^2+\sigma_w^2} < 1$. Now, $(\cos(\theta_2[k])\widetilde{w}_{2x}[k], \mathcal{F}_{k+2})$ is a Martingale Difference Sequence. Also, $v_x[k+1] \in \mathcal{F}_{k+1}$, since $v_x[k+1] = z_x[k+1] - z_x[k] - x[k+1] - x[k]$. So, the Martingale Stability

Theorem (MST) [49] applies, and we have

$$\sum_{\substack{k=0,\\ \cos(\theta_2[k])\neq 0}}^{t-1} \cos(\theta_2[k])\widetilde{w}_{2x}[k]v_x[k+1] = o(\sum_{\substack{k=0,\\ \cos(\theta_2[k])\neq 0}}^{t-1} v_x^2[k+1]) + O(1).$$

Substituting the above in (9.18), and the result in (9.15), we get

$$\sum_{\substack{k=0,\\ \cos(\theta_2[k])\neq 0}}^{t-1} v_x^2[k+1] + \frac{2h\beta}{1-\beta}\sum_{\substack{k=0,\\ \cos(\theta_2[k])\neq 0}}^{t-1} \cos(\theta_2[k])e_{2v}[k]v_x[k+1] + o(\sum_{\substack{k=0,\\ \cos(\theta_2[k])\neq 0}}^{t-1} v_x^2[k+1]) + O(1)$$
$$+ \sum_{\substack{k=0,\\ \cos(\theta_2[k])= 0}}^{t-1} v_x^2[k+1] = o(t). \qquad (9.19)$$

Dividing the above by t, taking the limit as $t \to \infty$, and invoking (9.17) completes the proof. □

Now we demonstrate the performance of dynamic watermarking for the transportation system. The specific strategy that the sensor uses to fabricate the measurements is the same as (9.4), except that now, $n[t] \sim \mathcal{N}(0, \sigma_x^2 + \sigma_e^2)$.. The two tests (9.8) and (9.9) of dynamic watermarking are employed by the controller. The asymptotic tests (9.8) and (9.9) were converted to statistical tests for implementation by checking if for each time, the LHS of (9.8) and (9.9) are within thresholds ϵ_1 and ϵ_2 respectively of their asymptotic values.

Fig. 9.3 plots the LHS of (9.8), and Fig. 9.4 plots that of (9.9) as a function of time. As can be seen, the false measurements pass test 2 but fail test 1, indicating an attack. The restoration of collision freedom for the automatic transportation system is shown in the attached video clip.

9.4 Concluding Remarks

In this chapter, we have addressed the problem of cybersecurity of advanced transportation systems, whether autonomous or those that provide driver assist. This is an exemplar of the broader class of cyber-physical systems for which there is great current concern on the issue of security. We have shown how collisions can be caused in such systems by attacking the sensor, even though the control logic contains a collision avoidance module. To provide security against such attacks, we

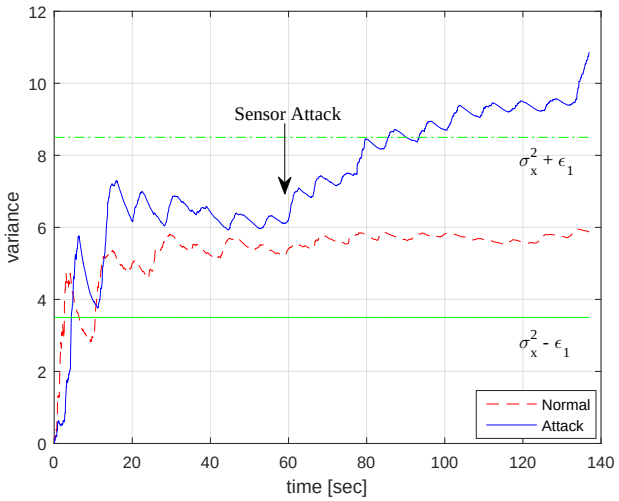

Figure 9.3: Test statistic of error test 1 as a function of time. Reprinted with permission from [2].

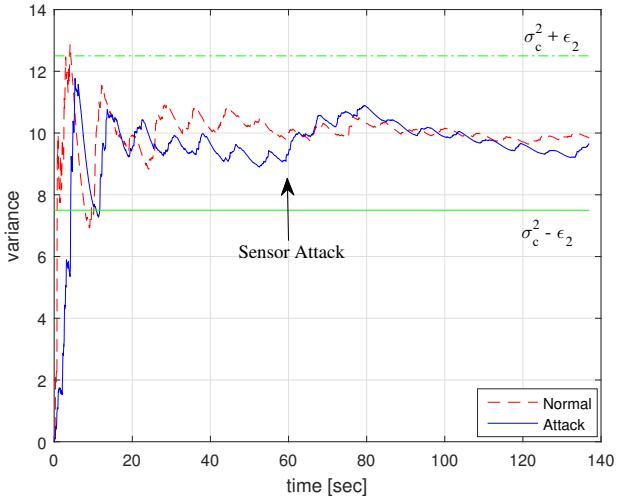

Figure 9.4: Test statistic of error test 2 as a function of time. Reprinted with permission from [2].

have considered the usage of dynamic watermarking where actuators inject small private excitation signals into the system and check the reported sensor measurements for statistical consistency with the injected noise. We have extended the theory of dynamic watermarking that has been developed for linear systems to nonlinear systems describing the equations of vehicular motion. We have implemented dynamic watermarking on a prototypical laboratory transportation systems and have shown how it restores collision freedom.

10. CONCLUSIONS

In this thesis, we have considered the problem of security of networked cyber-physical systems. Securing such systems is not the same as securing a communication network due to the fundamental role played by sensors and actuators in interfacing with and controlling the physical plant. In this thesis, we have studied certain fundamental problems arising in this context.

Analogous to the classical notions of controllable and unobservable subspaces of a linear system, we have introduced in Chapter 3 the notions of securable and unsecurable subspaces of a linear dynamical system, which have important operational meanings in the context of secure control. In Chapter 3, we have presented their operational meaning in the context of deterministic linear dynamical systems, and in Chapters 4 and 5, we have extended them to the context of stochastic linear dynamical systems. Specifically, we have shown that the malicious sensors and actuators cannot distort the state estimation performance of the honest nodes along the securable subspace of the system, no matter what attack strategy they employ; any performance degradation that they can cause is restricted to the unsecurable subspace of the system if they are to remain undetected.

We have developed in Chapter 6 the general procedure of dynamic watermarking that imposes private excitation signals on the actuation signals whose presence can then be traced around the loop to detect malicious behaviors of sensors. We have also rigorously established the security guarantees provided by dynamic watermarking in a variety of control system contexts, including a class of nonlinear systems. Chapter 7 addresses the problem of designing security-guaranteeing dynamic watermarks. Specifically, given a finite-dimensional perfectly-observed linear dynamical system affected by process noise of arbitrary probability distribution, we have derived in Chapter 7 the necessary and sufficient conditions that the probability distribution of the watermark should satisfy in order for the fundamental security guarantee of dynamic watermarking to hold. Chapter 8 addresses a scenario in which there are malicious sensors and actuators mixed with honest sensors and actuators, and wherein the honest actuators watermark their control inputs to detect malicious sensors. The state space of any such system can be decomposed into two orthogonal subspaces,

called the watermark-securable and the watermark-unsecurable subspaces, such that the malicious sensors and actuators cannot degrade the state estimation performance of the honest sensors and actuators along the watermark-securable subspace. The watermark-securable subspace is larger than the securable subspace that can be guaranteed in the absence of Dynamic Watermarking. A certain subset of the watermark-securable subspace was characterized in Chapter 8. Finally, in Chapter 9, we have demonstrated the efficacy of dynamic watermarking in securing a prototypical intelligent transportation system.

REFERENCES

[1] B. Satchidanandan and P. R. Kumar, "Dynamic watermarking: Active defense of networked cyber-physical systems," *Proceedings of the IEEE*, vol. 105, pp. 219–240, Feb 2017.

[2] Woo-Hyun Ko, B. Satchidanandan, and P. R. Kumar, "Theory and implementation of dynamic watermarking for cybersecurity of advanced transportation systems," in *2016 IEEE Conference on Communications and Network Security (CNS)*, pp. 416–420, Oct 2016.

[3] A. Cardenas, S. Amin, B. Sinopoli, A. Giani, A. Perrig, and S. Sastry, "Challenges for securing cyber physical systems," in *Workshop on future directions in cyber-physical systems security*, 2009.

[4] Y. Mo, T. H.-J. Kim, K. Brancik, D. Dickinson, H. Lee, A. Perrig, and B. Sinopoli, "Cyber-physical security of a smart grid infrastructure," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 195–209, 2012.

[5] M. Abrams and J. Weiss, "Malicious control system cyber security attack case study-Maroochy water services, Australia," *Technical Report*, 2008.

[6] G. Liang, S. R. Weller, J. Zhao, F. Luo, and Z. Y. Dong, "The 2015 ukraine blackout: Implications for false data injection attacks," *IEEE Transactions on Power Systems*, vol. 32, pp. 3317–3318, July 2017.

[7] R. Langner, "Stuxnet: Dissecting a cyberwarfare weapon," *Security & Privacy, IEEE*, vol. 9, no. 3, pp. 49–51, 2011.

[8] S. Staniford, V. Paxson, and N. Weaver, "How to own the internet in your spare time," in *Proceedings of the 11th USENIX Security Symposium*, (Berkeley, CA, USA), pp. 149–167, USENIX Association, 2002.

[9] R. Kalman, "On the general theory of control systems," in *IRE Transactions on Automatic Control*, 1959.

[10] J. T. Brassil, S. Low, N. F. Maxemchuk, and L. O. Gorman, "Electronic marking and identification techniques to discourage document copying," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 8, pp. 1495–1504, 1995.

[11] F. Hartung and M. Kutter, "Multimedia watermarking techniques," *Proceedings of the IEEE*, vol. 87, no. 7, pp. 1079–1107, 1999.

[12] A. Z. Tirkel, G. Rankin, R. Van Schyndel, W. Ho, N. Mee, and C. F. Osborne, "Electronic watermark," *Digital Image Computing, Technology and Applications (DICTAâĂŹ93)*, pp. 666–673, 1993.

[13] F. Pasqualetti, F. Dörfler, and F. Bullo, "Cyber-physical attacks in power networks: Models, fundamental limitations and monitor design," in *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on*, pp. 2195–2201, IEEE, 2011.

[14] F. Pasqualetti, F. Dorfler, and F. Bullo, "Control-theoretic methods for cyberphysical security: Geometric principles for optimal cross-layer resilient control systems," *IEEE Control Systems*, vol. 35, no. 1, pp. 110–127, 2015.

[15] F. Pasqualetti, F. Dörfler, and F. Bullo, "Cyber-physical security via geometric control: Distributed monitoring and malicious attacks," in *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, pp. 3418–3425, IEEE, 2012.

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

[17] "President's Commission on Critical Infrastructure Protection. Critical Foundations: Protecting America's Infrastructures." `https://www.fas.org/sgp/library/pccip.pdf`, 1997. Online; Accessed June 9, 2019.

[18] Y. Y. Haimes, N. C. Matalas, J. H. Lambert, B. A. Jackson, and J. F. R. Fellows, "Reducing vulnerability of water supply systems to attack," *Journal of Infrastructure Systems*, vol. 4, no. 4, pp. 164–177, 1998.

[19] M. Bishop and E. Goldman, "The strategy and tactics of information warfare," *Contemporary Security Policy*, vol. 24, no. 1, pp. 113–139, 2003.

[20] K. Barnes and B. Johnson, "Introduction to SCADA: Protection and vulnerabilities," *Technical Report INEEL/EXT-04-01710, Idaho National Engineering and Environmental Laboratory*, 2004.

[21] Y. Y. Haimes, "Risk of terrorism to cyber-physical and organizational-societal infrastructures," *Public Works Management and Policy*, vol. 6, no. 4, pp. 231–240, 2002.

[22] D. Watts, "Security and vulnerability in electric power systems," *35th North American Power Symposium*, pp. 559–566, 2003.

[23] A. S. Brown, "SCADA vs. the hackers," *Mechanical Engineering*, vol. 124, no. 12, p. 37, 2002.

[24] J. Eisenhauer, P. Donnelly, M. Ellis, and M. O'Brien, "Roadmap to secure control systems in the energy sector," *Energetics Incorporated. Sponsored by the U.S. Department of Energy and the U.S. Department of Homeland Security*, January 2006.

[25] "Roadmap to secure control systems in the water sector." `http://www.awwa.org/Portals/0/files/legreg/Security/SecurityRoadmap.pdf`, March 2006.

[26] "Roadmap to secure control systems in the chemical sector." `https://scadahacker.com/library/Documents/Roadmaps/Roadmap%20to%20Secure%20Control%20Systems%20in%20the%20Chemical%20Sector.pdf`, September 2009.

[27] "Roadmap to secure control systems in the transportation sector." `https://ics-cert.us-cert.gov/sites/default/files/ICSJWG-Archive/TransportationRoadmap20120831.pdf`, August 2012.

[28] A. A. Cardenas, S. Amin, and S. Sastry, "Secure control: Towards survivable cyber-physical systems," in *The 28th International Conference on Distributed Computing Systems Workshops*, IEEE, 2008.

[29] J. Ponniah, Y.-C. Hu, and P. R. Kumar, "A clean slate approach to secure wireless networking," *Foundations and Trends in Networking*, vol. 9, no. 1, pp. 1–105, 2014.

[30] A. A. Cardenas, S. Amin, and S. Sastry, "Research challenges for the security of control systems."

[31] L. Schenato, B. Sinopoli, M. Franceschetti, K. Poolla, and S. S. Sastry, "Foundations of control and estimation over lossy networks," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 163–187, 2007.

[32] B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, and S. Sastry, "Optimal linear LQG control over lossy networks without packet acknowledgment," *Asian Journal of Control*, vol. 10, no. 1, pp. 3–13, 2008.

[33] V. Gupta, B. Hassibi, and R. M. Murray, "Optimal LQG control across packet-dropping links," *Systems & Control Letters*, vol. 56, no. 6, pp. 439–446, 2007.

[34] N. Elia and S. K. Mitter, "Stabilization of linear systems with limited information," *IEEE Transactions on Automatic Control*, vol. 46, no. 9, pp. 1384–1400, 2001.

[35] X. Liu and A. Goldsmith, "Kalman filtering with partial observation losses," in *43rd IEEE Conference on Decision and Control*, vol. 4, IEEE, 2004.

[36] S. Amin, A. A. Cárdenas, and S. S. Sastry, "Safe and secure networked control systems under denial-of-service attacks," in *Hybrid Systems: Computation and Control*, Springer, 2009.

[37] Y. Mo, E. Garone, A. Casavola, and B. Sinopoli, "False data injection attacks against state estimation in wireless sensor networks," in *49th IEEE Conference on Decision and Control (CDC)*, IEEE, 2010.

[38] C.-Z. Bai, F. Pasqualetti, and V. Gupta, "Security in stochastic control systems: Fundamental limitations and performance bounds," *American Control Conference*, pp. 195âĂŞ–200, 2015.

[39] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure state-estimation for dynamical systems under active adversaries," in *49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE, 2011.

[40] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *47th Annual Allerton Conference on Communication, Control, and Computing*, Sept 2009.

[41] Y. Mo, R. Chabukswar, and B. Sinopoli, "Detecting integrity attacks on SCADA systems," *IEEE Transactions on Control Systems Technology*, vol. 22, no. 4, pp. 1396–1407, 2014.

[42] Y. Mo, S. Weerakkody, and B. Sinopoli, "Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs," *IEEE Control Systems*, vol. 35, pp. 93–109, Feb 2015.

[43] S. Weerakkody, Y. Mo, and B. Sinopoli, "Detecting integrity attacks on control systems using robust physical watermarking," in *53rd IEEE Conference on Decision and Control*, pp. 3757–3764, Dec 2014.

[44] F. Miao, M. Pajic, and G. J. Pappas, "Stochastic game approach for replay attack detection," in *2013 IEEE 52nd Annual Conference on Decision and Control (CDC)*, IEEE, 2013.

[45] A. Teixeira, I. Shames, H. Sandberg, and K. Johansson, "Revealing stealthy attacks in control systems," in *50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Oct 2012.

[46] S. Gisdakis, T. Giannetsos, and P. Papadimitratos, "Shield: A data verification framework for participatory sensing systems," in *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, WiSec '15, (New York, NY, USA), ACM, 2015.

[47] G. Basile and G. Marro, "Controlled and conditioned invariant subspaces in linear system theory," in *Journal of Optimization Theory and Applications*, 1969.

[48] B. Satchidanandan and P. R. Kumar, "Control systems under attack: The securable and unsecurable subspaces of a linear stochastic system," in *Emerging Applications of Control and Systems Theory*, pp. 217–228, Springer, 2018.

[49] T. L. Lai and C. Z. Wei, "Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems," *The Annals of Statistics*, pp. 154–166, 1982.

[50] B. Satchidanandan and P. R. Kumar, "The securable subspace of a linear stochastic system with malicious sensors and actuators," in *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 911–917, Oct 2017.

[51] J. Ponniah, Y. Hu, and P. R. Kumar, "A system-theoretic clean slate approach to provably secure ad-hoc wireless networking," *IEEE Transactions on Control of Network Systems*, vol. 3, pp. 206–217, June 2016.

[52] I.-H. Hou, V. Borkar, and P. R. Kumar, "A theory of qos for wireless," in *IEEE INFOCOM*, IEEE, 2009.

[53] R. Kalman, "On the general theory of control systems," *IRE Transactions on Automatic Control*, vol. 4, pp. 110–110, Dec 1959.

[54] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control*. SIAM Classics in Applied Mathematics, SIAM, Philadelphia, PA, USA, 2015.

[55] K. J. Åstrom, *Introduction to Stochastic Control*. New York: Academic Press, 1970, 1970.

[56] B. A. Francis and W. M. Wonham, "The internal model principle of control theory," *Automatica*, vol. 12, no. 5, pp. 457–465, 1976.

[57] T. Kailath, "The innovations approach to detection and estimation theory," *Proceedings of the IEEE*, vol. 58, no. 5, pp. 680–695, 1970.

[58] R. K. Mehra and J. Peschon, "An innovations approach to fault detection and diagnosis in dynamic systems," *Automatica*, vol. 7, no. 5, pp. 637–640, 1971.

[59] A. Wald, "Sequential tests of statistical hypotheses," *The Annals of Mathematical Statistics*, vol. 16, no. 2, pp. 117–186, 1945.

[60] T. W. Anderson, *An introduction to multivariate statistical analysis*, vol. 2. Wiley New York, 1958.

[61] M. Porter, A. Joshi, P. Hespanhol, A. Aswani, M. Johnson-Roberson, and R. Vasudevan, "Simulation and real-world evaluation of attack detection schemes," *arXiv preprint arXiv:1810.07773*, 2018.

[62] B. Satchidanandan and P. R. Kumar, "On minimal tests of sensor veracity for dynamic watermarking-based defense of cyber-physical systems," in *Communication Systems and Networks (COMSNETS), 2017 9th International Conference on*, pp. 23–30, IEEE, 2017.

[63] B. Satchidanandan and P. R. Kumar, "Secure control of networked cyber-physical systems," in *Decision and Control (CDC), 2016 IEEE 55th Conference on*, pp. 283–289, IEEE, 2016.

[64] A. Greenberg, "Hackers remotely kill a jeep on the highway- with me in it." `https://www.wired.com/2015/07/hackers-remotely-kill-jeep-highway/`. Online; accessed 9 June 2019.

[65] C. Miller and C. Valasek, "Remote exploitation of an unaltered passenger vehicle," *Black Hat, USA*, 2015.

[66] "Hackers reveal nasty new car attacks–with me behind the wheel (video)." `http://www.forbes.com/sites/andygreenberg/2013/07/24/hackers-reveal-nasty-new-car-attacks-with-me-behind-the-wheel-video/#29621f635bf2`. Online; accessed 9 June, 2019.

[67] J. Vanian, "Security experts say that hacking cars is easy." `http://fortune.com/2016/01/26/security-experts-hack-cars/`. Online; accessed 9 June, 2019.

[68] J. Kiss, "Your next car will be hacked. will autonomous vehicles be worth it?." `https://www.theguardian.com/technology/2016/mar/13/autonomous-cars-self-driving-hack-mikko-hypponen-sxsw`. Online; accessed 9 June, 2019.

[69] C. L. Robinson, H.-J. Schutz, G. Baliga, and P. R. Kumar, “Architecture and algorithm for a laboratory vehicle collision avoidance system,” in *2007 IEEE 22nd International Symposium on Intelligent Control*, pp. 23–28, IEEE, 2007.