

# Secure Control of Cyber-Physical Systems with Intermittent Data Authentication

by

Ilija M. Jovanov

Department of Electrical & Computer Engineering  
Duke University

Date: \_\_\_\_\_

Approved:

---

Miroslav Pajic, Supervisor

---

Michael Zavlanos

---

Krishnendu Chakrabarty

Thesis submitted in partial fulfillment of the requirements for the degree of  
Master of Science in the Department of Electrical & Computer Engineering  
in the Graduate School of Duke University  
2018

Copyright © 2018 by Ilija M. Jovanov  
All rights reserved except the rights granted by the  
Creative Commons Attribution-Noncommercial Licence

# Abstract

The increase in network connectivity has also resulted in several high-profile attacks on cyber-physical systems. An attacker that manages to access a local network could *remotely* affect control performance by tampering with sensor measurements delivered to the controller. Recent results have shown that with network-based attacks, such as *Man-in-the-Middle* attacks, the attacker can introduce an unbounded state estimation error if measurements from a suitable subset of sensors contain false data when delivered to the controller. While these attacks can be addressed with the standard cryptographic tools that ensure data integrity, their continuous use would introduce significant communication and computation overhead. Consequently, we study effects of intermittent data integrity guarantees on system performance under stealthy attacks. We consider linear estimators equipped with a general type of residual-based intrusion detectors (including  $\chi^2$  and CUSUM detectors), and show that even when integrity of sensor measurements is enforced only intermittently, the attack impact is significantly limited; specifically, the state estimation error is bounded or the attacker cannot remain stealthy. Furthermore, we present methods to: (1) evaluate the effects of any given integrity enforcement policy in terms of reachable state-estimation errors for any type of stealthy attacks, and (2) design an enforcement policy that provides the desired estimation error guarantees under attack. Finally, on three automotive case studies we show that even with less than 10% of authenticated messages we can ensure satisfiable control performance in the presence of attacks.

# Contents

<b>Abstract</b>	<b>iii</b>
<b>List of Figures</b>	<b>vi</b>
<b>List of Abbreviations and Symbols</b>	<b>ix</b>
<b>Acknowledgements</b>	<b>xiii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Notation and Terminology . . . . .	5
<b>2 Problem Description</b>	<b>7</b>
2.1 System Model without Attacks . . . . .	7
2.2 Attack Model . . . . .	9
2.3 Problem Formulation . . . . .	12
<b>3 Impact of Stealthy Attacks on State Estimation Error</b>	<b>14</b>
3.1 Perfectly Attackable Systems . . . . .	20
<b>4 Stealthy Attacks in Systems with Intermittent Integrity Enforcement</b>	<b>22</b>
4.0.1 Guarantees with Sensor-wise Integrity Enforcement . . . . .	32
<b>5 Analysis and Design of Safe Integrity Enforcement Policies</b>	<b>34</b>
5.1 $\chi^2$ detector . . . . .	34
5.1.1 Evaluation of the State Estimation Error Regions . . . . .	35
5.2 Reachable State Estimation Errors with Intermittent Integrity Enforce- ments (CUSUM) . . . . .	38

5.3	Design of Periodic Integrity Enforcement Policies (CUSUM) . . . . .	45
<b>6</b>	<b>Cumulative Message Authentication</b>	<b>48</b>
6.1	Reachability Analysis for Systems with Intermittent Cumulative Integrity Enforcements . . . . .	51
<b>7</b>	<b>Case Studies</b>	<b>56</b>
7.1	Case Study: Vehicle Trajectory Following . . . . .	56
7.2	Degraded Cooperative Adaptive Cruise Control (dCACC) . . . . .	58
7.3	Case Study: Delayed Authentication . . . . .	60
<b>8</b>	<b>Implementation</b>	<b>63</b>
8.1	Car and Platoon System Identification . . . . .	65
8.2	Estimation and Intrusion Detection . . . . .	68
	<b>Bibliography</b>	<b>70</b>

# List of Figures

1.1	Communication schedule for periodic messages (with period $T_s = 2$ ) from two sensors over a shared network: (a) a feasible schedule for non-authenticated messages (i.e., when MAC bits are not attached to transmitted packets); (b) there is no feasible schedule when all messages are authenticated; and (c) if data integrity is only intermittently enforced (e.g., by adding MAC bits only to every other packet), scheduling of the messages becomes feasible. . . . .	4
2.1	System architecture – by launching <i>Man-in-the-Middle</i> (MitM) attacks, the attacker can inject adversarial signals into plant measurements obtained from system sensors. . . . .	8
4.1	System evolution between two consecutive endpoints of integrity enforcement intervals. . . . .	28
5.1	4-step reachable regions for 2-norm and $\infty$ -norm. . . . .	38
6.1	Stealthy attacks performed by the attacker, depending on the threshold of the $\mathcal{R}_{risk}$ region. Cumulative MAC is assumed to arrive at $t_j = t + f_c$ . When Threshold 1 describes $\mathcal{R}_{risk}$ region, "Attack 1" is detected after cumulative MAC is received at $t_j$ without reaching $\mathcal{R}_{risk}$ , and thus attacker has to perform "Attack 2" to remain stealthy. On the other hand, when $\mathcal{R}_{risk}$ boundary is Threshold 2, the attacker reaches $\mathcal{R}_{risk}$ states before authentication, and successfully damages the system with "Attack 1" before it is detected. . . . .	50
7.1	Evolution of the maximal estimation error for vehicle tracking; without integrity enforcements, the attacker forces the system outside of the safe range in 4 steps. . . . .	57
7.2	Maximal estimation error in the presence of attacks on all sensors for vehicle-tracking case study with three different integrity enforcement policies with $f = 1$ and periods $L = 20, 30, 35$ . . . . .	58

7.3	State estimation of the tracked vehicle trajectory - without integrity enforcements a stealthy attacker can introduce a significant estimation error in a short period of time. However, even with intermittent integrity enforcement, the attack effects are negligible. Duration of the simulation is 200 s and the attack starts at 100 s. . . . .	58
7.4	Evolution of maximal estimation error for dCACC. If we can enforce integrity on two sensor values after every twenty unsecured sensor values, the system remains under the specified safety threshold. . . . .	60
7.5	Reachable state estimation errors in the presence of stealthy attacks for dCACC in steps $k = 11$ and $k = 22$ with and without data integrity enforcement. Without integrity enforcement, the size of reachable regions keeps increasing, while when integrity is being enforced with policy $L = 20$ and $f = 2$ , estimation error evolves as in Figure 7.4, and the attacker is contained between red and blue ellipsoids. . . . .	60
7.6	Evolution of the norm of estimation error. Figure shows two system realizations – one that uses unauthenticated data, and the other that authenticates previous $f = 4$ points of data every $L = 4$ samples. Authentication policy defined like this retains the estimation error inside of the safe states. . . . .	62
7.7	A 4s simulation of the system, with the attack gaining access to the sensors at 2s. Figure shows an error that the attacker could induce prior to being detected when authentication arrives. This figure shows possible attacks that depend on when the attacker injects erroneous values, and thus only one of the peaks shown in the figure could be reached prior to detection of the attack. . . . .	62
8.1	Snapshot of the car platoon system used for tests. . . . .	63
8.2	Full platooning system, with cars moving on top of a treadmill. ROS sends steering and speed commands to cars over the wireless network. Position and heading of cars is then recorded by a camera with help of AprilTags, and sent back to ROS through camera interface. ROS utilizes this data to compute corrected steering and speed that it will send to the cars. . . . .	64
8.3	Longitudinal movement controller for a vehicle in the vehicle platoon. Desired velocity in relation to camera is 0 once the goal position is achieved, and thus we include conveyor velocity. . . . .	65
8.4	Signal used for system identification along with the outputs of the system. Horizontal axis is provided in samples as subspace system identification returns system in discrete time. . . . .	66

8.5	Residuals obtained from one-car Kalman filter during the attacks. We can see residual spiking for a high-risk high-influence attack. . . . .	69
8.6	States of a single car. We can see that ramp attacks cause significant change in states without significant effect on residual. . . . .	69
8.7	Residuals obtained from whole platoon Kalman filter during the attacks. We can see residual spiking for a high-risk high-influence attack. . . . .	69
8.8	States of a three-car platoon. We can see that ramp attacks cause significant change in states without significant effect on residual. . . . .	69

# List of Abbreviations and Symbols

## Symbols

<b>A</b>	Matrix of state-space representation of the system, mapping states into future states.
<b>B</b>	Matrix that maps inputs into corresponding states.
<b>C</b>	Matrix that maps states into corresponding outputs.
<b>D</b>	Matrix that maps inputs into corresponding outputs.
$\mathbf{x}_k$	Vector of states at time $k$ .
$n$	Number of system states.
$\mathbf{u}_k$	Vector of inputs at time $k$ .
$m$	Number of system inputs.
$\mathbf{y}_k$	Vector of outputs at time $k$ .
$p$	Number of system outputs.
$\mathbf{w}_k$	Modeling (process) noise at time $k$ .
$\mathbf{v}_k$	Sensor (measuring) noise at time $k$ .
<b>R</b>	Sensor noise covariance matrix.
$\mathbb{R}$	Set of real numbers.
<b>K</b>	Estimator (Kalman) gain.
$\hat{\mathbf{x}}_k$	Vector of estimated states at time $k$ .
$\mathbf{z}_k$	Residual of the estimator at time $k$ .
<b>Q</b>	Covariance matrix of the residual.

$\mathbf{e}_k$	Estimation error at time $k$ .
$\Sigma$	Estimation error covariance matrix.
$g_k$	Value of intrusion detection function at time $k$ .
$c_i$	Weight coefficients in detection function sum.
$\mathcal{T}$	Length of the detection function window.
$\beta_k$	Probability of the alarm at time $k$ .
$h$	Decision threshold of the intrusion detector.
$\mathcal{S}$	Set of plant's sensors.
$\mathcal{K}$	Subset of plant's sensors.
$\mathbf{a}_k$	Signal injected by the attacker at time $k$ .
$\Delta \mathbf{e}_k$	Difference between compromised and non-compromised $\mathbf{e}_k$ .
$\Delta \mathbf{z}_k$	Difference between compromised and non-compromised $\mathbf{z}_k$ .
$\Xi$	Dynamical system describing evolution of $\Delta \mathbf{e}_k$ and $\Delta \mathbf{z}_k$ .
$\varepsilon$	Increase in probability of detection introduced by the attacker.
$\mathcal{A}_k$	Set of all stealthy attacks up to time $k$ .
$\mathcal{R}_k$	Set of all estimation errors reachable in $k$ steps.
$\mathcal{R}$	Global reachable region of the state estimation error.
$\alpha$	Robustness constraint corresponding to $\beta$ .
$\underline{\alpha}, \bar{\alpha}$	Underapproximation and overapproximation of $\alpha$ .
$\lambda$	Non-centrality parameter of non-central $\chi^2$ distribution.
$\hat{\mathcal{R}}_k^\alpha$	Set that has equal boundedness as set $\mathcal{R}_k$ .
$\lambda_i$	$i^{th}$ eigenvalue of a matrix.
$\mathbb{N}_0$	Set of all non-negative integers.
$\mathbf{v}$	Eigenvector of a matrix.
$\mu$	Sequence of time points when integrity is enforced.
$f$	Number of consecutive time points where integrity is enforces.

$L$	Maximal time distance between any two consecutive integrity enforcements.
$t_i$	$i^{th}$ element of the sequence $\mu$ .
$\mathbf{V}$	Matrix formed from eigenvectors of a matrix.
$\mathbf{J}$	Jordan matrix (jordan form of a matrix).
$\psi$	Observability index of the $(\mathbf{A}, \mathbf{C})$ pair.
$q_{un}$	Number of unstable eigenvalues of $\mathbf{A}$ .
$\mathcal{O}$	Observability matrix.
$\tilde{\mathcal{K}}_j$	Set of compromised sensor measurements at time $j$ .
$\mathcal{Q}_j$	Set of compromised sensor measurements up to time $j$ .
$\mathbf{P}$	Projection matrix.
$\mathbf{I}$	Identity matrix.
$\mathcal{H}_0$	Hypothesis that there is no attack present.
$\mathcal{H}_1$	Hypothesis that there is attack present.
$\Lambda_k$	log likelihood ratio at time $k$ .

## Abbreviations

CPS	Cyber-physical system.
SCADA	Supervisory control and data acquisition.
GPS	Global positioning system.
MitM	Man-in-the-Middle (attacks).
CUSUM	Cumulative sum (detector).
SPRT	Sequential probability ratio test.
MAC	Message authentication code.
CAN	Controller area network.
LTI	Linear time-invariant (system).
PA	Perfectly attackable.

ROS    Robot Operating System

# Acknowledgements

I would like to thank Intel and Duke University for financial support throughout my master studies. I would also like to thank professor Sebastian Fischmeister from University of Waterloo, who provided us with a vehicle platooning platform that allowed practical implementation of theoretical results presented in this thesis. Furthermore, I would like to thank two students from his lab, David Donghyun Shin and Mohammad Hossein Basiri, for their support during the work on vehicle platooning platform. I would also like to thank professor Henry P. Gavin for providing me with system identification knowledge and matlab codes needed for practical evaluation of this thesis. Last, but not the least, I would like to thank my advisor Miroslav Pajic for valuable direction in bringing this work to completion.

This work was supported in part by the NSF CNS-1652544 and CNS-1505701 grants, and the Intel-NSF Partnership for Cyber-Physical Systems Security and Privacy. This material is also based on research sponsored by the ONR under agreements number N00014-17-1-2012 and N00014-17-1-2504. Some of the preliminary results have appeared in [Jovanov and Pajic (2017b)] and [Jovanov and Pajic (2017a)].

## Introduction

Several high-profile incidents have recently exposed vulnerabilities of cyber-physical systems (CPS) and drawn attention to the challenges of providing security guarantees as part of their design. These incidents cover a wide range of application domains and system complexity, from attacks on large-scale infrastructure such as the 2016 breach of Ukrainian power-grid [Zetter (2016)], to the StuxNet virus attack on an industrial SCADA system [Langner (2011)], as well as attacks on controllers in modern cars (e.g., [Checkoway et al. (2011)]) and unmanned arial vehicles [Shepard et al. (2012)]

There are several reasons for such number of security related incidents affecting control of CPS. The tight interaction between information technology and physical world has greatly increased the attack vector space. For instance, an adversarial signal can be injected into measurements obtained from a sensor, using non-invasive attacks that modify the sensor's physical environment; as shown in attacks on GPS-based navigation systems [Tippenhauer et al. (2011); Kerns et al. (2014)]. Even more important reason is network connectivity that is prevalent in CPS. An attacker that manages to access a local control network could *remotely* affect control performance by tampering with sensor measurements and actuator commands in order to force the plant into any desired state,

as illustrated in [Smith (2011)]. From the controls perspective, attacks over an internal system network, such as the *Man-in-the-Middle* (MitM) attacks where the attacker inserts messages anywhere in the sensors→controllers→actuators pathway, can be modeled as additional malicious signals injected into the control loop via the system’s sensors and actuators [Teixeira et al. (2012)].

While the interaction with the physical world introduces new attack surfaces, it also provides opportunities to improve system resilience against attacks. The use of control techniques that employ a physical model of the system’s dynamics for attack detection and attack-resilient state estimation has drawn significant attention in recent years (e.g., [Teixeira et al. (2012, 2015); Pasqualetti et al. (2013); Fawzi et al. (2014); Sundaram et al. (2010); Mo et al. (2015); Amin et al. (2015); Pajic et al. (2014, 2017); Shoukry et al. (2016)], and a recent survey [Lun et al. (2016)]). One line of work is based on the use of unknown input observers (e.g., [Sundaram et al. (2010); Pasqualetti et al. (2013)]) and non-convex optimization for resilient estimation (e.g., [Fawzi et al. (2014); Pajic et al. (2017)]), while another focuses on attack-detection and estimation guarantees in systems with standard Kalman filter-based state estimators (e.g., [Mo and Sinopoli (2010); Mo et al. (2012); Kwon et al. (2014, 2015); Mo et al. (2015); Mo and Sinopoli (2016); Kwon and Hwang (2017)]). In the later works, estimation residue-based failure detectors, such as  $\chi^2$  [Mo and Sinopoli (2010, 2016)], cumulative sum (CUSUM) [R et al. (2017)], and sequential probability ratio test (SPRT) detectors [Kwon et al. (2015)], are employed for intrusion detection. Still, irrelevant of the utilized attack detection mechanism, after compromising a suitable subset of sensors, an intelligent attacker can significantly degrade control performance while remaining undetected (i.e., stealthy). For instance, for resilient state estimation techniques as in [Fawzi et al. (2014); Pajic et al. (2017)], measurements from at least half of the sensors should not be tampered with [Fawzi et al. (2014); Shoukry and Tabuada (2016)], while [Mo and Sinopoli (2010); Kwon et al. (2014)] capture attack requirements for Kalman filter-based estimators. The reason for such conservative results

lies in the common initial assumption that once a sensor or its communication to the estimator is compromised, all values received from the sensors can be potentially corrupted – i.e., integrity of the data received from these sensors cannot be guaranteed.

On the other hand, most of network-based attacks, including MitM attacks, can be avoided with the use of standard cryptographic tools. For example, to authenticate data and ensure integrity of received communication packets, a common approach is to add a message authentication code (MAC) to the transmitted sensor measurements. Therefore, data integrity requirements can be imposed by the continuous use of MACs in all transmissions from a sufficient subset of sensors. However, the overhead caused by the continuous computation and communication of authentication codes can limit their use. For instance, adding MAC bits to networked control systems that employ Controller Area Networks (CAN) may not be feasible due to the message length limitation (e.g., only 64 payload bits per packet in the basic CAN protocol), while splitting them into several communication packets significantly increases the message transmission time [Lin et al. (2015)]. To illustrate this, consider two sensors periodically transmitting measurements over a shared network. As presented in Figure 1.1(a), without authentication (i.e., if transmitted data contain no MAC bits) the communication packets will be schedulable but the system would be vulnerable to false-data injection attacks. Yet, if all measurements from both sensors are authenticated, with the increase in the packet size due to authentication overhead, it is not possible to schedule transmissions from both sensors in every communication frame (Figure 1.1(b)). Finally, a feasible schedule exists if MAC bits are attached to every other measurement packet transmitted by each sensor (Figure 1.1(c)).

Consequently, in this thesis we focus on state estimation in systems with *intermittent data integrity guarantees* for sensor measurements delivered to the estimator. Specifically, we study the performance of linear filters equipped with residual-based intrusion detectors in the presence of attacks on sensor measurements. We build on the system model from [Mo and Sinopoli (2010); Kwon et al. (2014); Mo and Sinopoli (2016)] by capturing



FIGURE 1.1: Communication schedule for periodic messages (with period  $T_s = 2$ ) from two sensors over a shared network: (a) a feasible schedule for non-authenticated messages (i.e., when MAC bits are not attached to transmitted packets); (b) there is no feasible schedule when all messages are authenticated; and (c) if data integrity is only intermittently enforced (e.g., by adding MAC bits only to every other packet), scheduling of the messages becomes feasible.

that the use of authentication mechanisms in intermittent time-points ensures that sensor measurements received at these points are valid. To keep our discussion and results general, we consider a wide class of detection functions that encompasses commonly used detectors, including  $\chi^2$  and CUSUM detectors. We show that even when integrity of communicated sensor data is enforced only intermittently and the attacker is fully aware of the times of the enforcement, the attack impact gets significantly limited; concretely, either the state estimation error remains bounded or the attacker cannot remain stealthy. This holds even when communication from *all* sensors to the estimator can be compromised as well as in any other case where otherwise (i.e., without integrity enforcements) an unbounded estimation error can be introduced.

Furthermore, to facilitate the use of intermittent data integrity enforcement for control of CPS in the presence of network-based attacks, we introduce an analysis and design framework that addresses two challenges. First, we introduce techniques to evaluate the effects of any given integrity enforcement policy in terms of reachable state-estimation errors for any type of stealthy attacks. Note that methods to evaluate potential state estimation errors due to attacks are considered in [Mo and Sinopoli (2016); Kwon et al. (2015); Mo and Sinopoli (2010)]. However, given that the previous work considers system architectures without intermittent use of authentication, these techniques result in overly

conservative estimates of reachable regions or they cannot capture the effects of intermittent integrity guarantees on the estimation error. Second, we present a method to design an enforcement policy that provides the desired estimation error guarantees for any attack signal inserted via compromised sensors. The developed framework also facilitates tradeoff analysis between the allowed estimation error and the rate at which data integrity should be enforced – i.e., the required system resources such as communication bandwidth as we have presented in [Lesi et al. (2017a)].

The rest of the thesis is organized as follows. In Chapter 2, we introduce the problem, including the system and attack models. In Chapter 3, we analyze the impact of stealthy attacks in systems without integrity enforcements and formally define intermittent integrity enforcement policies. Chapter 4 focuses on state estimation guarantees when data integrity is at least intermittently enforced. We then introduce a methodology to analyze effects of integrity enforcement policies as well as design suitable policies that ensure the desired estimation error even in the presence of attacks (Chapter 5). Finally, in Chapter 6 we present effects of delayed authentication, before presenting case studies that illustrate effectiveness of our approach in Chapter 7 and practical implementation of our approach in Chapter 8.

## 1.1 Notation and Terminology

The transpose of matrix  $\mathbf{A}$  is specified as  $\mathbf{A}^T$ , while the  $i^{th}$  element of a vector  $\mathbf{x}_k$  is denoted by  $\mathbf{x}_{k,i}$ . Moore-Penrose pseudoinverse of matrix  $\mathbf{A}$  is denoted as  $\mathbf{A}^\dagger$ . In addition,  $\|\mathbf{A}\|_i$  denotes the  $i$ -norm of a matrix  $\mathbf{A}$  and, for a positive definite matrix  $\mathbf{Q}$ ,  $\|\Delta\mathbf{z}_k\|_{\mathbf{Q}^{-1}} = \|\mathbf{Q}^{-1/2}\Delta\mathbf{z}_k\|_2$ .  $null(\mathbf{A})$  denotes the null space of the matrix. Also,  $\text{diag}(\cdot)$  indicates a square matrix with the quantities inside the brackets on the diagonal, and zeros elsewhere, while  $\text{BlckDiag}(\cdot)$  denotes a block-diagonal operator. We denote positive definite and positive semidefinite matrix  $\mathbf{A}$  as  $\mathbf{A} > 0$  and  $\mathbf{A} \geq 0$ , respectively, while  $\det(\mathbf{A})$  stands

for the determinant of the matrix. Also,  $\mathbf{I}_p$  denotes the  $p$ -dimensional identity matrix, and  $\mathbf{0}_{p \times q}$  denotes  $p \times q$  matrix of zeroes. We use  $\mathbb{R}, \mathbb{N}$  and  $\mathbb{N}_0$  to denote the sets of reals, natural numbers and nonnegative integers, respectively. As most of our analysis considers bounded-input systems, we refer to any eigenvalue  $\lambda$  as *unstable eigenvalue* if  $|\lambda| \geq 1$ .

For a set  $\mathcal{S}$ , we use  $|\mathcal{S}|$  to denote the cardinality (i.e., size) of the set. In addition, for a set  $\mathcal{K} \subset \mathcal{S}$ , with  $\mathcal{K}^c$  we denote the complement set of  $\mathcal{K}$  with respect to  $\mathcal{S}$  – i.e.,  $\mathcal{K}^c = \mathcal{S} \setminus \mathcal{K}$ . Projection vector  $\mathbf{i}_j^T$  denotes the row vector (of the appropriate size) with a 1 in its  $j^{\text{th}}$  position being the only nonzero element of the vector. For a vector  $\mathbf{y} \in \mathbb{R}^p$ , we use  $\mathbf{P}_{\mathcal{K}}\mathbf{y}$  to denote the projection from the set  $\mathcal{S} = \{1, \dots, p\}$  to set  $\mathcal{K}$  ( $\mathcal{K} \subseteq \mathcal{S}$ ) by keeping only elements of  $\mathbf{y}$  with indices from  $\mathcal{K}$ .<sup>1</sup> Finally, *the support* of the vector  $\mathbf{v} \in \mathbb{R}^p$  is the set  $\text{supp}(\mathbf{v}) = \{i \mid \mathbf{v}_i \neq 0\} \subseteq \{1, 2, \dots, p\}$ .

---

<sup>1</sup> Formally,  $\mathbf{P}_{\mathcal{K}} = [\mathbf{i}_{k_1} \dots \mathbf{i}_{k_{|\mathcal{K}|}}]^T$ , where  $\mathcal{K} = \{s_{k_1}, \dots, s_{k_{|\mathcal{K}|}}\} \subseteq \mathcal{S}$  and  $k_1 < k_2 < \dots < k_{|\mathcal{K}|}$ .

## Problem Description

Before introducing the problem formulation, we describe the considered system and its architecture (shown in Figure 2.1), as well as the attacker model.

### 2.1 System Model without Attacks

We consider an observable linear-time invariant (LTI) system whose evolution without attacks can be represented as

$$\begin{aligned}\mathbf{x}_{k+1} &= \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k + \mathbf{w}_k \\ \mathbf{y}_k &= \mathbf{C}\mathbf{x}_k + \mathbf{v}_k\end{aligned}\tag{2.1}$$

where  $\mathbf{x}_k \in \mathbb{R}^n$  and  $\mathbf{u}_k \in \mathbb{R}^m$  denote the plant's state and input vectors, at time  $k$ , while the plant's output vector  $\mathbf{y}_k \in \mathbb{R}^p$  contains measurements provided by  $p$  sensors from the set  $\mathcal{S} = \{s_1, s_2, \dots, s_p\}$ . Accordingly, the matrices  $\mathbf{A}$ ,  $\mathbf{B}$  and  $\mathbf{C}$  have suitable dimensions. Also,  $\mathbf{w} \in \mathbb{R}^n$  and  $\mathbf{v} \in \mathbb{R}^p$  denote the process and measurement noise; we assume that  $\mathbf{x}_0$ ,  $\mathbf{w}_k$ , and  $\mathbf{v}_k$  are independent Gaussian random variables.

Furthermore, the system is equipped with an estimator in the form of a Kalman filter. Given that the Kalman gain usually converges in only a few steps, to simplify the notation

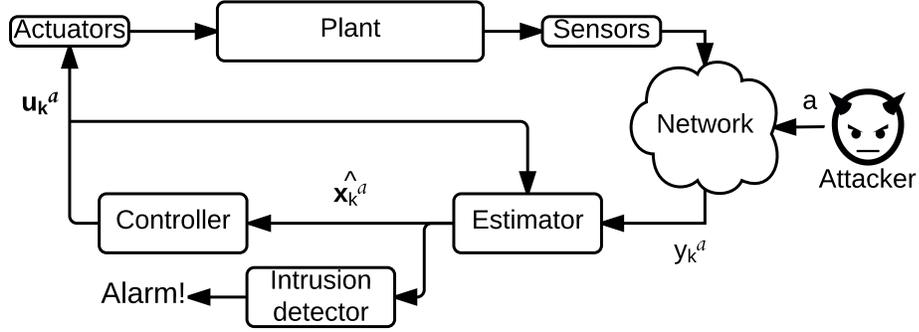


FIGURE 2.1: System architecture – by launching *Man-in-the-Middle* (MitM) attacks, the attacker can inject adversarial signals into plant measurements obtained from system sensors.

we assume that the system is in steady state before the attack. Hence, the Kalman filter estimate  $\hat{\mathbf{x}}_k$  is updated as

$$\hat{\mathbf{x}}_{k+1} = \mathbf{A}\hat{\mathbf{x}}_k + \mathbf{B}\mathbf{u}_k + \mathbf{K}(\mathbf{y}_{k+1} - \mathbf{C}(\mathbf{A}\hat{\mathbf{x}}_k + \mathbf{B}\mathbf{u}_k)) \quad (2.2)$$

$$\mathbf{K} = \mathbf{\Sigma}\mathbf{C}^T(\mathbf{C}\mathbf{\Sigma}\mathbf{C}^T + \mathbf{R})^{-1}, \quad (2.3)$$

where  $\mathbf{\Sigma}$  is the estimation error covariance matrix, and  $\mathbf{R}$  is the sensor noise covariance matrix. Also, the residue  $\mathbf{z}_k \in \mathbb{R}^p$  at time  $k$  and its covariance matrix  $\mathbf{Q}$  are defined as

$$\begin{aligned} \mathbf{z}_k &= \mathbf{y}_k - \mathbf{C}(\mathbf{A}\hat{\mathbf{x}}_{k-1} + \mathbf{B}\mathbf{u}_{k-1}), \\ \mathbf{Q} &= E\{\mathbf{z}_k\mathbf{z}_k^T\} = \mathbf{C}\mathbf{\Sigma}\mathbf{C}^T + \mathbf{R}. \end{aligned} \quad (2.4)$$

Finally, the state estimation error  $\mathbf{e}_k$  is defined as the difference between the plant's state  $\mathbf{x}_k$  and Kalman filter estimate  $\hat{\mathbf{x}}_k$  as

$$\mathbf{e}_k = \mathbf{x}_k - \hat{\mathbf{x}}_k. \quad (2.5)$$

In addition to the estimator, we assume that the system is equipped with an intrusion detector. We consider a general case where the detection function  $g_k$  of the intrusion detector is defined as

$$g_k = \sum_{i=k-\mathcal{T}+1}^k c_{(i-k+\mathcal{T})} \mathbf{z}_i^T \mathbf{Q}^{-1} \mathbf{z}_i. \quad (2.6)$$

Here,  $\mathcal{T}$  is the length of the detector's time window, and  $c_i$  for  $i = 1, \dots, \mathcal{T}$  are predefined non-negative coefficients, with  $c_{\mathcal{T}}$  being strictly positive. The above formulation captures both fixed window size detectors, where  $\mathcal{T}$  is a constant, as well as detectors where the time window size  $\mathcal{T}$  satisfies  $\mathcal{T} = k$ . Also, the definition of the detection function  $g_k$  covers a wide variety of commonly used intrusion detectors, such as  $\chi^2$  and cumulative sum (CUSUM) detectors previously considered in these scenarios [Mo et al. (2010); Miao et al. (2013); Mo and Sinopoli (2016); Miao et al. (2017); Kwon et al. (2014); R et al. (2017)]. The alarm is triggered when the value of the detection function  $g_k$  satisfies

$$g_k > \text{threshold}, \quad (2.7)$$

and the probability of the alarm at time  $k$  can be captured as

$$\beta_k = P(g_k > \text{threshold}). \quad (2.8)$$

## 2.2 Attack Model

We assume that the attacker is capable of launching MitM attacks on communication channels between a subset of the plant's sensors  $\mathcal{K} \subseteq \mathcal{S}$  and the estimator; for instance, by secretly relaying corresponding altered communication packets. However, we do not assume that the set  $\mathcal{K}$  is known to the system or system designers. Thus, to capture the attacker's impact on the system, the system model from (2.1) becomes

$$\begin{aligned} \mathbf{x}_{k+1}^a &= \mathbf{A}\mathbf{x}_k^a + \mathbf{B}\mathbf{u}_k^a + \mathbf{w}_k \\ \mathbf{y}_k^a &= \mathbf{C}\mathbf{x}_k^a + \mathbf{v}_k + \mathbf{a}_k. \end{aligned} \quad (2.9)$$

Here,  $\mathbf{x}_k^a$  and  $\mathbf{y}_k^a$  denote the state and plant outputs in the presence of attacks, from the perspective of the estimator, since in the general case they differ from the plant's state and outputs of the non-compromised system. In addition,  $\mathbf{a}_k$  denotes the signals injected by the attacker at time  $k$  starting from  $k = 1$  (i.e.,  $\mathbf{a}_0 = \mathbf{0}$ );<sup>1</sup> to model MitM attacks on

<sup>1</sup> More details about why the attacker does not insert attack at step  $k = 0$  can be found in Remark 1.

communication between the sensors from set  $\mathcal{K}$  and the estimator, we assume that  $\mathbf{a}_k$  is a sparse vector from  $\mathbb{R}^p$  with support in the set  $\mathcal{K}$  – i.e.,  $\mathbf{a}_{k,i} = 0$  for all  $i \in \mathcal{K}^C$  and  $k > 0$ .<sup>2</sup>

We consider the following *threat model*.

(1) The attacker has full knowledge of the system – in addition to knowing the dynamical model of the plant, employed Kalman filter, and detector, the attacker is aware of all potential security mechanism used in communication. Specifically, we consider systems that use standard methods for message authentication to ensure data integrity, and assume that the attacker is aware at which time points data integrity will be enforced. Thus, to avoid being detected, the attacker will not launch attacks in these steps and will also take into account these integrity enforcements in planning its attacks (as described in Chapter 3).<sup>3</sup> Since we model our system such that attacks start at  $k = 1$ , this further implies that at  $k = 1$  data integrity is not enforced, as otherwise the attacker would not be able to insert false data.

(2) The attacker has the required computation power to calculate suitable attack signals, while planning ahead as needed. (S)he also has the ability to inject *any* signal using communication packets mimicking sensors from the set  $\mathcal{K}$ , except at times when data integrity is enforced. For instance, when MACs are used to ensure data integrity and authenticity of communication packets, our assumption is that the attacker does not know the shared secret key used to generate the MACs.

*The goal of the attacker* is to design attack signal  $\mathbf{a}_k$  such that it *maximizes the error of state estimation* while ensuring that *the attack remains stealthy*. To formally capture this objective and the stealthiness constraint, we denote the state estimation, residue, and estimation error of the compromised system by  $\hat{\mathbf{x}}_k^a$ ,  $\mathbf{z}_k^a$ , and  $\mathbf{e}_k^a$ , respectively. Thus, the

---

<sup>2</sup> Although a sensor itself may not be directly compromised with MitM attacks, but rather communication between the sensor and estimator, we will also refer to these sensors are *compromised sensors*. In addition, in this work we sometimes abuse the notation by using  $\mathcal{K}$  to denote both the set of compromised sensors and the set of indices of the compromised sensors.

<sup>3</sup> In Chapter 4, we will also consider the case where the attacker has limited knowledge of the system's use of security mechanisms.

attacker's aim is to maximize  $e_k^a$ , while ensuring that the increase in the probability of alarm is not significant. We also define as

$$\Delta \mathbf{e}_k = \mathbf{e}_k^a - \mathbf{e}_k, \quad \Delta \mathbf{z}_k = \mathbf{z}_k^a - \mathbf{z}_k,$$

the change in the estimation error and residue, respectively, caused by the attacks. From (2.1) and (2.9), the evolution of these signals can be captured as a dynamical system  $\Xi$  of the form

$$\Delta \mathbf{e}_{k+1} = (\mathbf{A} - \mathbf{KCA})\Delta \mathbf{e}_k - \mathbf{Ka}_{k+1}, \quad (2.10)$$

$$\Delta \mathbf{z}_k = \mathbf{CA}\Delta \mathbf{e}_{k-1} + \mathbf{a}_k, \quad (2.11)$$

with  $\Delta \mathbf{e}_0 = \mathbf{0}$ .

**Remark 1.** *From the above equations, the first attack vector to affect the change in estimation error is  $\mathbf{a}_1$ . Thus, without loss of generality, we assume that the attack starts at  $k = 1$  (i.e.,  $\mathbf{a}_i = \mathbf{0}$ , for all  $i \leq 0$ ). This also implies that  $\Delta \mathbf{z}_0 = \mathbf{0}$ .*

Note that the above dynamical system is noiseless (and deterministic), with input  $\mathbf{a}_k$  controlled by the attacker. Therefore, since  $E[\mathbf{e}_k] = \mathbf{0}$  for the non-compromised system in steady state, it follows that

$$\Delta \mathbf{e}_k = E[\Delta \mathbf{e}_k] = E[\mathbf{e}_k^a]. \quad (2.12)$$

Given that  $\Delta \mathbf{e}_k$  provides expectation of the state estimation error under the attack, this signal can be used to evaluate the impact that the attacker has on the system.<sup>4</sup> Thus, we specify the objective of the attacker as to **maximize the expected state estimation error** (e.g.,  $\|\Delta \mathbf{e}_k\|_2$ ). This is additionally justified by the fact that since  $\mathbf{a}_k$  is controlled by the attacker (i.e., deterministic to simplify of our presentation), which implies

$$\text{Cov}(\mathbf{e}_k^a) = \text{Cov}(\mathbf{e}_k) = \Sigma. \quad (2.13)$$

---

<sup>4</sup> For this reason, and to simplify our presentation, in the rest of the thesis we will sometimes refer to  $\Delta \mathbf{e}_k$  as *the (expected) state estimation error* instead of *the change of the state estimation error* caused by attacks.

To capture the attacker’s stealthiness requirements, we use the probability of alarm in the presence of an attack

$$\beta_k^a = P(g_k^a > \text{threshold}), \quad \text{where} \quad (2.14)$$

$$g_k^a = \sum_{i=k-\mathcal{T}+1}^k c_{(i-k+\mathcal{T})} \mathbf{z}_i^{aT} \mathbf{Q}^{-1} \mathbf{z}_i^a. \quad (2.15)$$

Therefore, to ensure that attacks remain stealthy, the attacker’s *stealthiness constraint* in each step  $k$  is to maintain

$$\beta_k^a \leq \beta_k + \varepsilon, \quad (2.16)$$

for a small predefined value of  $\varepsilon > 0$ .

### 2.3 Problem Formulation

As we will present in the next chapter, for a large class of systems, a stealthy attacker can easily introduce an unbounded state estimation error by compromising communication between some of the sensors and the estimator. On the other hand, existing communication protocols commonly incorporate security mechanisms (e.g., MAC) that can ensure integrity of delivered sensor measurements. Specifically, this means that the system could enforce  $\mathbf{a}_{k,i} = 0$  for some sensor  $s_i$ , or  $\mathbf{a}_k = \mathbf{0}$  if integrity for all transmitted sensor measurements is enforced at some time-step  $k$ . However, as we previously described, the integrity enforcement comes at additional communication and computation cost, effectively preventing their continuous use in resource constrained CPS.

Consequently, we focus on the problem of evaluating the impact of stealthy attacks in systems with intermittent (i.e., occasional) use of data integrity enforcement mechanisms.<sup>5</sup> Specifically, we will address the following problems:

- Can the attacker introduce unbounded state estimation errors in systems with intermittent integrity guarantees?

<sup>5</sup> Formal definition of such policies are presented in the next chapter.

- How to efficiently evaluate the impact of intermittent integrity enforcement policies on the induced state estimation errors in the presence of a stealthy attacker?
- How to design a non-overly conservative development framework that incorporates guarantees for estimation degradation under attacks into design of suitable integrity enforcement policies?

## Impact of Stealthy Attacks on State Estimation Error

To capture the impact of stealthy attacks on the system, we start with the following definition.

**Definition 1.** *The set of all stealthy attacks up to time  $k$  is*

$$\mathcal{A}_k = \{\mathbf{a}_{1..k} | \beta_{k'}^a \leq \beta_{k'} + \varepsilon, \forall k', 1 \leq k' \leq k\}, \quad (3.1)$$

where  $\mathbf{a}_{1..k} = [\mathbf{a}_1^T \dots \mathbf{a}_k^T]^T$ .

When reasoning about a set of reachable state estimation errors  $\mathbf{e}_k^a$  due to stealthy attacks from  $\mathcal{A}_k$ , we have to also take into account the variability of the estimation error. From (2.13), we can define a specific region that will contain the error  $\mathbf{e}_k^a$  with a desired probability. Therefore, we introduce the following definition.

**Definition 2.** *The  $k$ -reachable region  $\mathcal{R}_k$  of the state estimation error under the attack (i.e.,  $\mathbf{e}_k^a$ ) is the set*

$$\mathcal{R}_k = \left\{ \mathbf{e} \in \mathbb{R}^n \mid \begin{array}{l} \mathbf{e}\mathbf{e}^T \leq E[\mathbf{e}_k^a]E[\mathbf{e}_k^a]^T + \gamma \text{Cov}(\mathbf{e}_k^a), \\ \mathbf{e}_k^a = \mathbf{e}_k^a(\mathbf{a}_{1..k}), \mathbf{a}_{1..k} \in \mathcal{A}_k \end{array} \right\}. \quad (3.2)$$

Furthermore, the global reachable region  $\mathcal{R}$  of the state estimation error  $\mathbf{e}_k^a$  is the set

$$\mathcal{R} = \bigcup_{k=1}^{\infty} \mathcal{R}_k. \quad (3.3)$$

Here,  $\gamma$  is a design parameter directly related to the desired confidence that  $\mathbf{e}_k^a$  belongs to the reachable region. Effectively, the set  $\mathcal{R}_k$  captures the set of state estimation errors that can be reached in  $k^{\text{th}}$  step due to the injected malicious signal, while  $\mathcal{R}$  captures the set of all reachable state estimation errors. To assess vulnerability of the system, a critical characteristic of  $\mathcal{R}$  is *boundedness* – whether a stealthy attacker can introduce unbounded estimation errors. To simplify the boundedness analysis of  $\mathcal{R}$ , we start with the following theorem.

**Theorem 1.** *Let  $g_k = \mathbf{z}_k^T \mathbf{Q}^{-1} \mathbf{z}_k$  be the detector function. Then, for any  $\varepsilon > 0$ , such that  $\varepsilon \leq 1 - \beta_k$ , there exists a unique  $\alpha > 0$  such that  $\beta_k^a \leq \beta_k + \varepsilon$  if and only if  $\|\Delta \mathbf{z}_k\|_{\mathbf{Q}^{-1}} \leq \alpha$ .*

*Proof.* In the case without attacks, in steady-state  $g_k$  has  $\chi^2$  distribution with  $p$  degrees of freedom, since the residue  $\mathbf{z}_k$  is zero-mean ( $E[\mathbf{z}_k] = \mathbf{0}$ ) with covariance matrix  $\mathbf{Q} = \mathbf{C}\Sigma\mathbf{C}^T + \mathbf{R}$  [Juang (1994); Johnson et al. (1995)]. Furthermore, from (2.10) and (2.11),  $\Delta \mathbf{z}_k = \mathbf{z}_k^a - \mathbf{z}_k$  is output of a deterministic system controlled by noiseless input  $\mathbf{a}_{1..k}$ , as this is the bias injected by the attacker over the network, making it independent from particular values of  $\mathbf{z}_k$ . Thus,  $\mathbf{z}_k^a$  is a non-zero mean with covariance matrix  $\mathbf{Q}$  – i.e., the attacker is only influencing the  $\Delta \mathbf{z}_k = E[\mathbf{z}_k^a - \mathbf{z}_k] = E[\mathbf{z}_k^a]$ . Therefore,  $\mathbf{g}_k^a = \mathbf{z}_k^{aT} \mathbf{Q}^{-1} \mathbf{z}_k^a$  will have a non-central  $\chi^2$  distribution with  $p$  degrees of freedom; the non-centrality parameter of this distribution will be  $\lambda = \|\Delta \mathbf{z}_k\|_{\mathbf{Q}^{-1}}^2$  [Johnson et al. (1995)].

Let  $h$  be the threshold for the detector in (2.7). The alarm probabilities  $\beta_k = 1 - P(g_k \leq h)$  and  $\beta_k^a = 1 - P(g_k^a \leq h)$  can be computed from the distributions for  $g_k$  and  $g_k^a$  as

$$\beta_k = 1 - F_{\chi^2}(h, p), \quad \beta_k^a = 1 - F_{nc\chi^2}(h; p, \lambda),$$

where  $F_{\chi^2}(h, p)$  and  $F_{nc\chi^2}(h; p, \lambda)$  are cumulative distribution functions of  $\chi^2$  and non-centralized  $\chi^2$ , respectfully, at  $h$ , with  $p$  degrees of freedom and noncentrality parameter  $\lambda$ . Since  $p$  and  $h$  are fixed by the system design, it follows that  $\beta_k$  will be a constant, and  $\beta_k^a$  will be a function of  $\lambda$ .

Consider  $\varepsilon = \beta_k^a - \beta_k$ . This means that

$$\varepsilon = 1 - F_{nc\chi^2}(h; p, \lambda) - \beta_k. \quad (3.4)$$

The probability distribution function of non-central  $\chi^2$  distribution is smooth (thus making  $F_{nc\chi^2}(h; p, \lambda)$  smooth), and  $F_{nc\chi^2}(h; p, \lambda)$  is a decreasing function of  $\lambda$  [Johnson et al. (1995)]. Hence, it follows that for any  $\varepsilon$  there will exist exactly one  $\sqrt{\lambda} = \|\Delta\mathbf{z}_k\|_{\mathbf{Q}^{-1}} = \alpha$  such that (3.4) is satisfied. Furthermore, for any  $\varepsilon'$  that is lower than  $\varepsilon$ , the corresponding  $\sqrt{\lambda'} = \|\Delta\mathbf{z}_k\|_{\mathbf{Q}^{-1}}$  from (3.4) has to be lower than  $\alpha$ , and vice versa, which concludes the proof.  $\square$

Since the bound  $\alpha$  for  $\|\Delta\mathbf{z}_k\|_{\mathbf{Q}^{-1}}$  in Theorem 1 depends on  $\varepsilon$ ,  $h$  and the fact that the  $\chi^2$  detector with  $p$  degrees of freedom is used, we will denote such value as  $\alpha = \alpha_{\chi^2}(\varepsilon, p, h)$ .

**Remark 2.** *Related results from [Mo et al. (2010); Mo and Sinopoli (2016)], focus only on the detection function  $g_k = \mathbf{z}_k^T \mathbf{Q}^{-1} \mathbf{z}_k$  and show only sufficient conditions for stealthy attacks – i.e., that in this case from a robustness condition  $\|\Delta\mathbf{z}_k\|_{\mathbf{Q}^{-1}} \leq \alpha$  it follows that the the stealthiness condition  $\beta_k^a \leq \beta + \varepsilon$  is satisfied. However, the equivalence between conditions  $\|\Delta\mathbf{z}_k\|_{\mathbf{Q}^{-1}} \leq \alpha$  and  $\beta_k^a \leq \beta + \varepsilon$  will enable us to reduce conservativeness of our analysis as well as analyze boundness of the reachability region for the general type of detection functions from (2.15), by allowing us to employ both conditions interchangeably.*

From Definition 1 and Theorem 1 the following result holds.

**Corollary 2.** *For the detection function  $g_k = \mathbf{z}_k^T \mathbf{Q}^{-1} \mathbf{z}_k$ , there exists  $\alpha > 0$  such that the set of all stealthy attacks satisfies*

$$\mathcal{A}_k = \{\mathbf{a}_{1..k} \mid \|\Delta\mathbf{z}_{k'}\|_{\mathbf{Q}^{-1}} \leq \alpha, \forall k', 1 \leq k' \leq k\}. \quad (3.5)$$

The previous results introduce an equivalent ‘robustness-based’ representation for the set of stealthy attacks in systems where  $\chi^2$  detectors are used. They also provide a foundation to consider the more general formulation (2.15) for the detector function. We start with the following results characterizing over- and under-approximations of the set  $\mathcal{A}_k$  in such case, also using suitable ‘robustness-based’ representations of the stealthiness condition. By showing that reachable estimation error regions are bounded for these sets of attacks, we will be able to reason whether the reachable region of state estimation errors is bounded for attacks from the set  $\mathcal{A}_k$ .

**Lemma 3.** *For a system with the detector function  $g_k^a$  of the form from (2.15), the set of all stealthy attacks  $\mathcal{A}_k$  can be underapproximated by the set*

$$\underline{\mathcal{A}}_k = \{\mathbf{a}_{1..k} \mid \|\Delta \mathbf{z}_{k'}\|_{\mathbf{Q}^{-1}} \leq \underline{\alpha}, \forall k', 1 \leq k' \leq k\} \quad (3.6)$$

(i.e.,  $\underline{\mathcal{A}}_k \subseteq \mathcal{A}_k$ ), where  $\underline{\alpha} = \alpha_{\chi^2}(\varepsilon, \mathcal{T}p, h/c_{max})/\sqrt{\mathcal{T}}$ .

In essence, the lemma states that if  $\|\Delta \mathbf{z}_k\|_{\mathbf{Q}^{-1}} \leq \underline{\alpha}$  holds, then  $g_k^a \leq h$  for the general detection function from (2.15) is satisfied with probability that is lower than or equal to  $\beta_k + \varepsilon$ .

*Proof.* Consider an attack sequence  $\mathbf{a}_{1..k} \in \underline{\mathcal{A}}_k$  and the resulting evolution of the system from (2.10) and (2.11), with  $\|\Delta \mathbf{z}_{k'}\|_{\mathbf{Q}^{-1}} \leq \underline{\alpha}$ , for all  $k', 1 \leq k' \leq k$ . Then,

$$\sum_{i=k-\mathcal{T}+1}^k \|\Delta \mathbf{z}_i\|_{\mathbf{Q}^{-1}}^2 \leq \sum_{i=k-\mathcal{T}+1}^k \underline{\alpha}^2 = \alpha_{\chi^2}^2(\varepsilon, \mathcal{T}p, h). \quad (3.7)$$

In addition, we define  $c_{max} = \max(c_1, \dots, c_{\mathcal{T}})$  and

$$\underline{g}_k^a = \sum_{i=k-\mathcal{T}+1}^k c_{max} \mathbf{z}_i^{aT} \mathbf{Q}^{-1} \mathbf{z}_i^a, \quad (3.8)$$

as well as  $\underline{\beta}_k^a = P(\underline{g}_k^a > h)$ . From (3.8),  $\underline{g}_k^a$  is a scaled sum of noncentral  $\chi^2$  distributions with  $p$  degrees of freedom, so  $\underline{g}_k^a/c_{max}$  will have the noncentral  $\chi^2$  distribution with  $p\mathcal{T}$  degrees of freedom and the central moment equal to

$$\underline{\lambda} = \sum_{i=k-\mathcal{T}+1}^k \|\Delta \mathbf{z}_i\|_{\mathbf{Q}^{-1}}^2. \quad (3.9)$$

Since  $\underline{\beta}_k^a = P(\underline{g}_k^a > h) = P(\underline{g}_k^a/c_{max} > h/c_{max})$ , following the proof of Theorem 1 for the  $\underline{g}_k^a/c_{max}$  detection function we have that  $\underline{\beta}_k^a \leq \beta_k + \varepsilon$  is satisfied if and only if  $\sqrt{\underline{\lambda}} \leq \alpha_{\chi^2}(\varepsilon, p\mathcal{T}, h/c_{max})$ . That is, using (3.9)

$$\left(\underline{\beta}_k^a \leq \beta_k + \varepsilon\right) \Leftrightarrow \sum_{i=k-\mathcal{T}+1}^k \|\Delta \mathbf{z}_i\|_{\mathbf{Q}^{-1}}^2 \leq \alpha_{\chi^2}^2(\varepsilon, p\mathcal{T}, h/c_{max}). \quad (3.10)$$

Since (3.7) follows from the condition of the theorem, from (3.10) we have that  $\underline{\beta}_k^a \leq \beta_k + \varepsilon$  is satisfied. From (3.8) we have that  $g_k^a \leq \underline{g}_k^a$ , meaning that  $\beta_k^a \leq \underline{\beta}_k^a$ . Thus,  $(\beta_k^a \leq \beta_k + \varepsilon)$  holds, and  $\mathbf{a}_{1..k} \in \mathcal{A}_k$  (i.e.,  $\underline{\mathcal{A}}_k \subseteq \mathcal{A}_k$ ).  $\square$

**Lemma 4.** *For a system with the detector function  $g_k^a$  of the form from (2.15), the set of all stealthy attacks at time  $k$ ,  $\mathcal{A}_k$ , can be overapproximated by the set*

$$\overline{\mathcal{A}}_k = \{\mathbf{a}_{1..k} \mid \|\Delta \mathbf{z}_{k'}\|_{\mathbf{Q}^{-1}} \leq \bar{\alpha}, \forall k', 1 \leq k' \leq k\}, \quad (3.11)$$

(i.e.,  $\overline{\mathcal{A}}_k \supseteq \mathcal{A}_k$ ), where  $\bar{\alpha} = \alpha_{\chi^2}(\varepsilon, p, h/c_{\mathcal{T}})$ .

*Proof.* Consider an attack sequence  $\mathbf{a}_{1..k} \in \mathcal{A}_k$  with the detector function  $g_k^a$  from (2.15). Let  $\overline{g}_k^a = c_{\mathcal{T}} \mathbf{z}_k^{aT} \mathbf{Q}^{-1} \mathbf{z}_k^a$ . Since  $\overline{g}_k^a \leq g_k^a$  it follows that  $\overline{\beta}_k^a = P(\overline{g}_k^a > h) \leq \beta_k^a$ , where  $\beta_k^a$  is defined as in (2.14). Since  $\mathbf{a}_{1..k}$  are stealthy, it follows that  $\beta_k^a \leq \beta_k + \varepsilon$ , and thus  $\overline{\beta}_k^a \leq \beta_k + \varepsilon$  holds.

On the other hand, the function  $\overline{g}_k^a/c_{\mathcal{T}}$  has the  $\chi^2$  distribution; by following the proof of Theorem 1 for  $\overline{g}_k^a/c_{\mathcal{T}}$  we have that  $\overline{\beta}_k^a \leq \beta_k + \varepsilon$  is satisfied if and only if  $\|\Delta \mathbf{z}_k\|_{\mathbf{Q}^{-1}} \leq$

$\bar{\alpha} = \alpha_{\chi^2}(\varepsilon, p, h/c_{\mathcal{T}})$ . Therefore, we have that  $\beta_{k'}^a \leq \beta_{k'} + \varepsilon$  implies  $\|\Delta \mathbf{z}_{k'}\|_{\mathbf{Q}^{-1}} \leq \bar{\alpha}$ ,  $k' = 1, \dots, k$ , meaning that  $\mathbf{a}_{1..k} \in \bar{\mathcal{A}}_k$  (i.e.,  $\mathcal{A}_k \subseteq \bar{\mathcal{A}}_k$ ).  $\square$

**Remark 3.** *The previous lemmas also hold for the detection function  $g_k = \sum_{i=1}^k c_k \mathbf{z}_i^T \mathbf{Q}^{-1} \mathbf{z}_i$ ; this can be shown by replacing  $\mathcal{T}$  with  $k$  in the previous analysis, since it would not affect their proofs. In essence, this means that these results hold for both windowed detectors and CUSUM detectors – CUSUM detectors are explored in detail in Chapter 5.*

Lemmas 3 and 4 introduce attack sets  $\underline{\mathcal{A}}_k$  and  $\bar{\mathcal{A}}_k$  for which the attack constraints are captured as robustness bounds on  $\|\Delta \mathbf{z}_k\|_{\mathbf{Q}^{-1}}$  instead of probabilities of attack detection, and for which  $\underline{\mathcal{A}}_k \subseteq \mathcal{A}_k \subseteq \bar{\mathcal{A}}_k$ . Hence, to analyze impact of stealthy attacks, we can consider the effects of attacks that have to maintain  $\|\Delta \mathbf{z}_k\|$  below a certain threshold.

**Theorem 5.**  $\mathcal{R}_k$  from (3.2) is bounded if and only if the set

$$\hat{\mathcal{R}}_k^\alpha = \left\{ \Delta \mathbf{e}_k \in \mathbb{R}^n \mid \begin{array}{l} \Delta \mathbf{e}_k, \Delta \mathbf{z}_k \text{ satisfy (2.10) and (2.11),} \\ \Delta \mathbf{e}_{k-1} \in \hat{\mathcal{R}}_{k-1}^\alpha, \|\Delta \mathbf{z}_k\|_2 \leq \alpha \end{array} \right\}, \quad (3.12)$$

is bounded, where  $\hat{\mathcal{R}}_0^\alpha = \mathbf{0} \in \mathbb{R}^n$  and  $\alpha > 0$ .

*Proof.* From (2.13),  $\gamma \text{Cov}(\mathbf{e}_k^a)$  is bounded and we can simplify our presentation by focusing on the case where  $\gamma = 0$ . Furthermore, for any vector  $\mathbf{v}$ , the set  $\{ \mathbf{e} \in \mathbb{R}^n \mid \mathbf{e} \mathbf{e}^T \leq \mathbf{v} \mathbf{v}^T \}$

is bounded if and only if the vector  $\mathbf{v}$  is bounded. Therefore, the set  $\{ \mathbf{e} \in \mathbb{R}^n \mid \mathbf{e} \mathbf{e}^T \leq E[\mathbf{e}_k^a] E[\mathbf{e}_k^a]^T + \gamma C$

will be bounded if and only if  $E[\mathbf{e}_k^a] = \Delta \mathbf{e}_k$  (from (2.12)) is bounded.

Consider attack vectors  $\mathbf{a}_{1..k} \in \mathcal{A}_k$ . From Lemmas 3 and 4 we have that

$$\{\Delta \mathbf{e}_k \mid \mathbf{a}_{1..k} \in \underline{\mathcal{A}}_k\} \subseteq \{\Delta \mathbf{e}_k \mid \mathbf{a}_{1..k} \in \mathcal{A}_k\} \subseteq \{\Delta \mathbf{e}_k \mid \mathbf{a}_{1..k} \in \bar{\mathcal{A}}_k\}, \quad (3.13)$$

where we somewhat abuse the notation, by having  $\{\Delta \mathbf{e}_k \mid \mathbf{a}_{1..k} \in \mathcal{A}\}$  capture all reachable vectors  $\Delta \mathbf{e}_k$  when the system (2.10) is ‘driven’ by attack vectors from the set  $\mathcal{A}$ .

On the other hand, from linearity of the system described by (2.10) and (2.11), the sets

$\{\Delta \mathbf{e}_k \|\Delta \mathbf{z}_{k'}\|_{\mathbf{Q}^{-1}} \leq \underline{\alpha}, k' = 1, \dots, k\}$  and  $\{\Delta \mathbf{e}_k \|\Delta \mathbf{z}_{k'}\|_{\mathbf{Q}^{-1}} \leq \bar{\alpha}, k' = 1, \dots, k\}$  are either both bounded or both unbounded. Thus, from (3.13), these sets are bounded if and only if  $\{\Delta \mathbf{e}_k | \mathbf{a}_{1..k} \in \mathcal{A}_k\}$  is bounded.

Finally, as  $\frac{1}{|\lambda_{max}|} \|\Delta \mathbf{z}_k\|_2 \leq \|\Delta \mathbf{z}_k\|_{\mathbf{Q}^{-1}} \leq \frac{1}{|\lambda_{min}|} \|\Delta \mathbf{z}_k\|_2$ , where  $\lambda_{max}, \lambda_{min}$  are the largest and smallest, respectively, eigenvalue of  $\mathbf{Q}$ , the region  $\hat{\mathcal{R}}_k^\alpha$  will be bounded for the constraint  $\|\Delta \mathbf{z}_k\|_{\mathbf{Q}^{-1}} \leq \alpha$  if and only if its bounded with a 2-norm stealthiness constraint  $\|\Delta \mathbf{z}_k\|_2 \leq \alpha$  from (3.12).  $\square$

### 3.1 Perfectly Attackable Systems

Theorem 5 can be used to formally capture dynamical systems for which there exists a stealthy attack sequence that results in an unbounded state estimation error – i.e., for such systems, given enough time, the attacker can make arbitrary changes in the system states without risking detection.

**Definition 3.** *A system is perfectly attackable (PA) if the system's reachable set  $\mathcal{R}$  from (3.3) is an unbounded set.*

As shown in [Mo and Sinopoli (2010); Kwon et al. (2014)], for LTI systems without any additional data integrity guarantees, the set  $\hat{\mathcal{R}}^\alpha = \bigcup_{k=0}^{\infty} \hat{\mathcal{R}}_k^\alpha$  can be bounded or unbounded depending on the system dynamics and the set of compromised sensors  $\mathcal{K}$ . From Theorem 5, this property is preserved for the set  $\mathcal{R}$  as well. For this reason, we will be using the definition of  $\hat{\mathcal{R}}^\alpha$  to analyze boundedness of  $\mathcal{R}$ , and to simplify the notation due to linearity of the constraint we will assume that  $\alpha = 1$  – i.e., for this analysis we consider the stealthiness attack constraint as

$$\|\Delta \mathbf{z}_k\|_2 \leq 1, \quad k \in \mathbb{N}_0, \quad (3.14)$$

imposed on the system  $\Xi$  from (2.10) and (2.11).

Now, the theorem below follows from [Mo and Sinopoli (2010); Kwon et al. (2014)].

**Theorem 6.** *A system from (2.9) is perfectly attackable if and only if the matrix  $\mathbf{A}$  is unstable, and at least one eigenvector  $\mathbf{v}$  corresponding to an unstable eigenvalue satisfies  $\text{supp}(\mathbf{C}\mathbf{v}) \subseteq \mathcal{K}$  and  $\mathbf{v}$  is a reachable state of the system  $\Xi$  from (2.10), (2.11).*

Note that [Mo and Sinopoli (2010)] also uses the term *unstable eigenvalue*  $\lambda$  to denote  $|\lambda| \geq 1$ . In the next chapter, we show that intermittent integrity guarantees significantly limit stealthy attacks even for *perfectly-attackable* systems.

## Stealthy Attacks in Systems with Intermittent Integrity Enforcement

In this chapter, we analyze the effects that intermittent data integrity guarantees have on the estimation error under attack. To formalize this notion, we start with the following definition.

**Definition 4.** A global *intermittent data integrity enforcement policy*  $(\mu, f, L)$ , where  $\mu = \{t_k\}_{k=0}^{\infty}$  such that  $t_0 > 1$ , for all  $k > 0$ ,  $t_{k-1} < t_k$  and  $L = \sup_{k>0} (t_k - t_{k-1})$ , ensures that

$$\mathbf{a}_{t_k} = \mathbf{a}_{t_k+1} = \dots = \mathbf{a}_{t_k+f-1} = \mathbf{0}, \forall k \geq 0.$$

Furthermore, for a sensor  $s_i \in \mathcal{S}$ , the *sensor's intermittent data integrity enforcement policy*  $(\mu_i, f_i, L_i)$ , where  $\mu_i = \{t_k^i\}_{k=0}^{\infty}$  with  $t_0^i > 1$ ,  $t_{k-1}^i < t_k^i$  for all  $k > 0$ , and  $L_i = \sup_{k>0} (t_k^i - t_{k-1}^i)$ , ensures that

$$\mathbf{a}_{t_k^i, i} = \mathbf{a}_{t_k^i+1, i} = \dots = \mathbf{a}_{t_k^i+f_i-1, i} = \mathbf{0}, \forall k \geq 0.$$

Intuitively, an intermittent data integrity enforcement policy for sensor  $s_i$  ensures that the injected attack  $\mathbf{a}_{k,i}$  via the sensor will be equal to zero in at least  $f_i$  consecutive points,

where the starts of these ‘blocks’ are at most  $L_i$  time-steps apart. Similarly, for a *global* intermittent data integrity enforcement policy, the whole attack vector  $\mathbf{a}_k$  has to be  $\mathbf{0}$  for at least  $f$  consecutive steps, and the duration between these blocks is bounded from above to at most  $L$  time-steps.

Global intermittent integrity enforcement is easier to model (and analyze, as we will show in the next chapter). However, compared to the use of  $p$  separate sensor’s intermittent integrity enforcements, global enforcement policies impose significantly larger communication and computation overhead in every time-step when data integrity is enforced. For example, with global enforcement every sensor has to be able to compute and add a MAC to its message transmitted over a shared bus during one sampling period (which usually corresponds to a single communication frame). In addition, since in these systems estimation/control updates are commonly computed once all messages are received, when the integrity is enforced the estimator has to be able to evaluate/recompute all received MACs before its execution for that time-period. On the other hand, with integrity enforcement for each sensor, their MACs can be sent and reevaluated in separate (e.g., consecutive) sampling periods (i.e., communication frames).

**Remark 4.** *It is worth noting that our definition of intermittent integrity enforcement policies imposes a maximum time between integrity enforcements which, as we will show, is related to the worst-case estimation error caused by the attacks. The definition also captures **periodic** integrity enforcements when  $L = t_k - t_{k-1}$  for all  $k > 0$ . Finally, the definition also allows for capturing policies with continuous integrity enforcements, by specifying  $L \leq f$ .*

The following theorem specifies that when a global intermittent integrity enforcement policy is used a stealthy attacker cannot insert an unbounded expected state estimation error.

**Theorem 7.** Consider an LTI system from (2.1) with a global data integrity policy  $(\mu, f, L)$ , where

$$f = \min(\psi, q_{un}), \quad (4.1)$$

$L$  is finite,  $\psi$  is the observability index of the  $(\mathbf{A}, \mathbf{C})$  pair, and  $q_{un}$  denotes the number of unstable eigenvalues of  $\mathbf{A}$ . Then the system is not perfectly attackable.

From the above theorem, it follows that even intermittent integrity guarantees significantly limit the damage that the attacker could make to the system. Furthermore, the theorem makes no assumptions about the set  $\mathcal{K}$  of compromised sensors; in the general case, system designers may not be able to provide this type of guarantees during system design, and thus no restrictions are imposed on the set, neither regarding the number of elements or whether some sensors belong to it.

**Remark 5.** In our preliminary results reported in [Jovanov and Pajic (2017b)], a similar formulation of Theorem 7 is used with  $f = \min(\text{nullity}(\mathbf{C}) + 1, q_{un})$ . Since  $\psi \leq n - \text{rank}(\mathbf{C}) + 1$  from [Sundaram (2012)], using the rank–nullity theorem it follows that  $\psi \leq \text{nullity}(\mathbf{C}) + 1$ , meaning that the condition from Theorem 7 is stronger than our earlier result and may further reduce the number of integrity-enforcement points.

In the rest of the thesis, we use the notation from Theorem 7 for  $f$  and  $q_{un}$ . To show the theorem, we exploit the following Lemma 8 and Theorem 9; the lemma states that if stealthy attacks introduce unbounded estimation error  $\Delta \mathbf{e}_k$ , the unbounded components must belong to vector subspaces corresponding to unstable modes of the system (i.e., matrix  $\mathbf{A}$ ).

**Lemma 8.** Consider system  $\Xi$  from (2.10) and (2.11) under the stealthiness constraint (3.14), and let us denote by  $\mathbf{v}_1, \dots, \mathbf{v}_{q_{un}}$  eigenvectors and generalized eigenvectors that correspond to unstable eigenvalues of matrix  $\mathbf{A}$ . Then, unbounded estimation errors  $\Delta \mathbf{e}_k$  can be represented as

$$\Delta \mathbf{e}_k = \alpha_1 \mathbf{v}_1 + \dots + \alpha_{q_{un}} \mathbf{v}_{q_{un}} + \varrho_k \quad (4.2)$$

where  $\varrho_k = \sum_{j=q_{un}+1}^n \alpha_j \mathbf{v}_j$  is a bounded vector, and for some  $1 \leq i \leq q_{un}$  it holds that  $\alpha_i \rightarrow \infty$  as  $k \rightarrow \infty$ .

*Proof.* The proof is provided in the appendix.  $\square$

**Theorem 9.** Consider any  $k \in \mathbb{N}$ , such that  $k + 1 \in \mu$  (i.e., at time  $k + 1$  an integrity enforcement block in the policy  $\mu$  starts). If  $\Delta \mathbf{e}_k$  is reachable state of  $\Xi$ , and if vectors  $\mathbf{CA}\Delta \mathbf{e}_k, \mathbf{CA}\Delta \mathbf{e}_{k+1}, \dots, \mathbf{CA}\Delta \mathbf{e}_{k+f-1}$  are bounded, then the vector  $\Delta \mathbf{e}_{k+f}$  has to be bounded for any stealthy attack.<sup>1</sup>

*Proof.* From (2.10) and (2.11) it follows that

$$\Delta \mathbf{e}_{k+f} = \mathbf{A}\Delta \mathbf{e}_{k+f-1} - \mathbf{K}\Delta \mathbf{z}_{k+f}, \quad (4.3)$$

$$\Delta \mathbf{z}_{k+f} = \mathbf{CA}\Delta \mathbf{e}_{k+f-1} + \mathbf{a}_{k+f}. \quad (4.4)$$

Since  $\|\mathbf{K}\|_2$  is bounded, and  $\|\Delta \mathbf{z}_{k+f}\|_2 \leq 1$  due to the stealthy attack constraint (3.14), then  $\|\mathbf{K}\Delta \mathbf{z}_{k+f}\|_2 \leq \|\mathbf{K}\|_2 \|\Delta \mathbf{z}_{k+f}\|_2$  is bounded. Thus, to show that  $\|\Delta \mathbf{e}_{k+f}\|_2$  is bounded, it is sufficient to prove that  $\|\mathbf{A}\Delta \mathbf{e}_{k+f-1}\|_2$  is bounded.

Let's assume the opposite – i.e., that  $\|\mathbf{A}\Delta \mathbf{e}_{k+f-1}\|_2$  is unbounded while

$$\|\mathbf{CA}\Delta \mathbf{e}_k\|_2, \dots, \|\mathbf{CA}\Delta \mathbf{e}_{k+f-1}\|_2$$

are all bounded. From (4.3) it follows that

$$\mathbf{A}\Delta \mathbf{e}_{k+f-1} = \mathbf{A}^f \Delta \mathbf{e}_k - \sum_{j=1}^{f-1} \mathbf{A}^{f-j} \mathbf{K}\Delta \mathbf{z}_{k+j}.$$

Given that  $\|\Delta \mathbf{z}_{k+1}\|_2, \dots, \|\Delta \mathbf{z}_{k+f-1}\|_2$  are bounded due to the stealthy attack requirements, in order for  $\mathbf{A}\Delta \mathbf{e}_{k+f-1}$  to be unbounded,  $\mathbf{A}^f \Delta \mathbf{e}_k$  has to be unbounded as well.

---

<sup>1</sup> Formally, the theorem states that the subsequence  $\{\Delta \mathbf{e}_{k+f}\}_{(k+1) \in \mu}$  of the sequence  $\{\Delta \mathbf{e}_k\}_{k \in \mathbb{N}}$  is bounded, if the subsequence  $\{\mathbf{CA}\Delta \mathbf{e}_k, \mathbf{CA}\Delta \mathbf{e}_{k+1}, \dots, \mathbf{CA}\Delta \mathbf{e}_{k+f-1}\}_{(k+1) \in \mu}$  of the sequence  $\{\mathbf{CA}\Delta \mathbf{e}_k\}_{k \in \mathbb{N}}$  is bounded. However, to simplify our presentation and notation, we simply refer to the vectors, instead of subsequences, as bounded.

Since  $\mathbf{CA}\Delta\mathbf{e}_{k+1}$  is bounded, this implies that  $\mathbf{CA}(\mathbf{A}\Delta\mathbf{e}_k - \mathbf{K}\Delta\mathbf{z}_{k+1})$  has to be bounded too. However, as  $\Delta\mathbf{z}_{k+1}$  has to be bounded due to the stealthiness condition, it follows that  $\mathbf{CA}^2\Delta\mathbf{e}_k$  has to remain bounded. Similarly, we can show that this holds up to  $\mathbf{CA}^f\Delta\mathbf{e}_k$ , and thus the vector  $\mathbf{b}_k(f)$  defined as

$$\mathbf{b}_k(f) = \underbrace{\begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \dots \\ \mathbf{CA}^{f-1} \end{bmatrix}}_{\mathcal{O}_f} \mathbf{A}\Delta\mathbf{e}_k \quad (4.5)$$

is bounded. Now, we consider two cases.

**Case I:** If  $f$  is observability index of  $(\mathbf{A}, \mathbf{C})$  pair (i.e.,  $f = \psi$ ), then  $\mathcal{O}_f$  has full rank, from which it follows that  $\mathbf{A}\Delta\mathbf{e}_k$  (and thus  $\mathbf{A}^f\Delta\mathbf{e}_k$ ) has to be also bounded, which is a contradiction.

**Case II:** Consider  $f = q_{un}$ , and let us use similarity transform  $\mathbf{V}$  on the initial system, where  $\mathbf{V}$  is defined as in the Lemma 8 proof – i.e.,  $\mathbf{V} = [\mathbf{v}_1 \dots \mathbf{v}_n]$  and we index (generalized) eigenvectors such that for each eigenvector  $\mathbf{v}_i$  with  $L_i$  generalized eigenvectors,  $\mathbf{v}_{i+1}, \dots, \mathbf{v}_{i+L_i}$  is its generalized eigenvector chain; in addition,  $\mathbf{v}_1, \dots, \mathbf{v}_{q_{un}}$  are the eigenvectors (including generalized eigenvectors) for all unstable modes of  $\mathbf{A}$ .

Thus, the transformed system can be captured as

$$\begin{aligned} \tilde{\mathbf{A}} &= \mathbf{V}^{-1}\mathbf{A}\mathbf{V} = \mathbf{J} = \begin{bmatrix} \mathbf{J}_1(q_{un} \times q_{un}) & \mathbf{0}_{(q_{un} \times n - q_{un})} \\ \mathbf{0}_{(n - q_{un} \times q_{un})} & \mathbf{J}_2(n - q_{un} \times n - q_{un}) \end{bmatrix} \\ \tilde{\mathbf{C}} &= \mathbf{C}\mathbf{V} = \begin{bmatrix} \tilde{\mathbf{C}}_1(p \times q_{un}) & \tilde{\mathbf{C}}_2(p \times n - q_{un}) \end{bmatrix} \end{aligned} \quad (4.6)$$

where  $\mathbf{J}$  is the Jordan form of  $\mathbf{A}$ ,  $\mathbf{J}_1$  captures unstable modes of  $\mathbf{A}$  and the pair  $(\tilde{\mathbf{A}}, \tilde{\mathbf{C}})$  is also observable.

Since  $\mathbf{A}^f\Delta\mathbf{e}_k$  is unbounded we have that  $\Delta\mathbf{e}_k$  is unbounded (from  $\|\mathbf{A}^f\Delta\mathbf{e}_k\|_2 \leq$

$\|\mathbf{A}\|_2^f \|\Delta \mathbf{e}_k\|_2$ ). Thus, from Lemma 8,

$$\Delta \mathbf{e}_k = \mathbf{V} [\alpha_1 \dots \alpha_n]^T = \mathbf{V} \alpha_{1..n}, \quad (4.7)$$

where  $\alpha_{1..q_{un}} = [\alpha_1 \dots \alpha_{q_{un}}]^T$  is unbounded while  $\alpha_{(q_{un}+1)..n} = [\alpha_{(q_{un}+1)} \dots \alpha_{q_n}]^T$  is a bounded vector. Due to the fact that  $\tilde{\mathbf{C}}\tilde{\mathbf{A}}^j = \tilde{\mathbf{C}}\mathbf{J}^j = \begin{bmatrix} \tilde{\mathbf{C}}_1 \mathbf{J}_1^j & \tilde{\mathbf{C}}_2 \mathbf{J}_2^j \end{bmatrix}$ , from (4.5) it follows that

$$\begin{aligned} \mathbf{b}_k(f) &= \begin{bmatrix} \tilde{\mathbf{C}}_1 & \tilde{\mathbf{C}}_2 \\ \tilde{\mathbf{C}}_1 \mathbf{J}_1 & \tilde{\mathbf{C}}_2 \mathbf{J}_2 \\ \dots & \dots \\ \tilde{\mathbf{C}}_1 \mathbf{J}_1^{f-1} & \tilde{\mathbf{C}}_2 \mathbf{J}_2^{f-1} \end{bmatrix} \mathbf{J} \alpha_{1..n} = \\ &= \underbrace{\begin{bmatrix} \tilde{\mathbf{C}}_1 \\ \tilde{\mathbf{C}}_1 \mathbf{J}_1 \\ \dots \\ \tilde{\mathbf{C}}_1 \mathbf{J}_1^{f-1} \end{bmatrix}}_{\tilde{\mathcal{O}}_{uns,f}} \mathbf{J}_1 \alpha_{1..q_{un}} + \underbrace{\begin{bmatrix} \tilde{\mathbf{C}}_2 \\ \tilde{\mathbf{C}}_2 \mathbf{J}_2 \\ \dots \\ \tilde{\mathbf{C}}_2 \mathbf{J}_2^{f-1} \end{bmatrix}}_{\tilde{\mathcal{O}}_{sta,f}} \mathbf{J}_2 \alpha_{(q_{un}+1)..n}. \end{aligned} \quad (4.8)$$

Since  $\mathbf{b}_k(f)$  and  $\alpha_{(q_{un}+1)..n}$  are bounded, from (4.8) the vector

$$\tilde{\mathbf{b}}_k(f) = \tilde{\mathcal{O}}_{uns,f} \mathbf{J}_1 \alpha_{1..q_{un}} \quad (4.9)$$

is also bounded. Note that  $\tilde{\mathcal{O}}_{uns,f}$  is effectively the observability matrix of the  $(\mathbf{J}_1, \tilde{\mathbf{C}}_1)$  pair corresponding to the subsystem with the  $q_{un}$  unstable eigenvalues of  $\mathbf{A}$ .

To show that  $(\mathbf{J}_1, \tilde{\mathbf{C}}_1)$  is observable, let us assume the opposite; thus, there exist an eigenvector  $\tilde{\mathbf{v}}_j$  of  $\mathbf{J}_1$  such that  $\tilde{\mathbf{v}}_j \in \text{null}(\tilde{\mathbf{C}}_1) \subseteq \mathbb{R}^{q_{un}}$ . Take note that  $\mathbf{J}_1$  is a Jordan matrix, so each of its eigenvectors has to be a projection vector  $\mathbf{i}_j$  (as defined in Sec. 1.1), where  $j$ ,  $1 \leq j \leq q_{un}$ , corresponds to the start of a Jordan block of  $\mathbf{J}_1$ . Yet this implies that  $\tilde{\mathbf{C}}_1 \mathbf{i}_j = \mathbf{0}_{p \times 1}$  - i.e., the  $j^{\text{th}}$  column of  $\tilde{\mathbf{C}}_1$  and thus the  $j^{\text{th}}$  column of  $\tilde{\mathbf{C}}$  are zero vectors. However, since  $\tilde{\mathbf{C}} = \mathbf{C}\mathbf{V} = [\mathbf{C}\mathbf{v}_1 \dots \mathbf{C}\mathbf{v}_n]$ , it follows that  $\mathbf{C}\mathbf{v}_j = \mathbf{0}$  for some  $j$ . Due to the way  $\mathbf{V}$  is formed and since  $j$  has to be the start index of a Jordan block in  $\mathbf{J}_1$ , it follows that  $\mathbf{v}_j$  is an eigenvector of  $\mathbf{A}$ . However, this implies that  $\mathbf{v}_j \in \text{null}(\mathbf{C})$ , making  $(\mathbf{A}, \mathbf{C})$  pair unobservable and contradicting our initial assumption about the system.

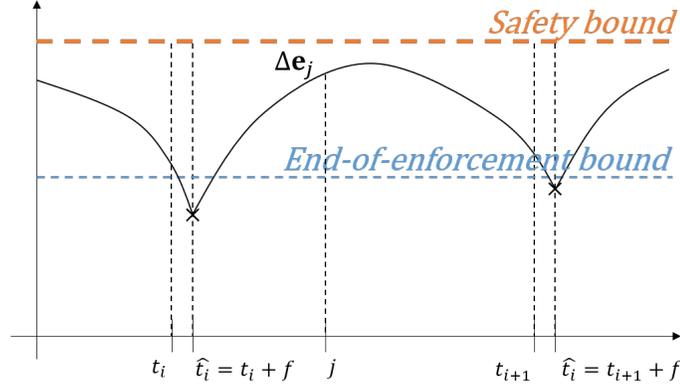


FIGURE 4.1: System evolution between two consecutive endpoints of integrity enforcement intervals.

Therefore,  $(\mathbf{J}_1, \tilde{\mathbf{C}}_1)$  is observable meaning that  $\tilde{\mathcal{O}}_{uns,f}$  is full rank. Furthermore,  $\mathbf{J}_1$  is invertible as it contains only unstable (i.e., non-zero) eigenvalues of the system on the diagonal. Hence, from (4.9) and the fact that  $\mathbf{b}_k(f)$  is bounded it follows that vector  $\alpha_{1..q_{un}}$  has to be bounded, which from (4.7) contradicts that  $\Delta \mathbf{e}_k$  and  $\mathbf{A}^f \Delta \mathbf{e}_k$  are unbounded, and thus concludes the proof.  $\square$

Using the previous theorem, we now prove Theorem 7.

*Proof of Theorem 7.* Consider any time-point  $t_k + f$  such that  $t_k \in \mu$  – i.e.,  $t_k$  is the start of an integrity enforcement block. Thus,  $\mathbf{a}_{t_k} = \dots = \mathbf{a}_{t_k+f-1} = \mathbf{0}$ . From (2.11) it follows that  $\Delta \mathbf{z}_{t_k+j} = \mathbf{CA} \Delta \mathbf{e}_{t_k+j-1}$ ,  $j = 0, \dots, f-1$ , and thus from (3.14)

$$\|\mathbf{CA} \Delta \mathbf{e}_{t_k+j-1}\|_2 \leq 1, \quad j = 0, \dots, f-1.$$

Now, from Theorem 9 it follows that the state estimation error  $\Delta \mathbf{e}_{t_k+f-1}$  has to be bounded for any stealthy attack; this holds for all time points at the ends of integrity enforcement intervals. Since in the proof of Theorem 9, we have not used any specifics of the time points, there exists a global bound on state estimation error at the end of all integrity enforcement periods (as illustrated in Figure 4.1).

Finally, consider an expected state-estimation error vector at any time  $j$ . From Definition 4, there exists  $t_i \in \mu$  such that  $j \in [\hat{t}_i, \hat{t}_i + L)$ , where  $\hat{t}_i = t_i + f$  (Figure 4.1). Now, from (2.10) and (2.11) we have that

$$\Delta \mathbf{e}_j = \mathbf{A}^{j-\hat{t}_i} \Delta \mathbf{e}_{\hat{t}_i} - \sum_{l=1}^{j-\hat{t}_i} \mathbf{A}^{j-\hat{t}_i-l} \mathbf{K} \Delta \mathbf{z}_{\hat{t}_i+l}. \quad (4.10)$$

Thus, the evolution of the expected state estimation error vector between two time points with bounded values can be described as evolution over a finite number of steps of a dynamical system with bounded inputs (since  $\|\Delta \mathbf{z}_{\hat{t}_i+l}\|_2 \leq 1$ ); from the triangle and Cauchy-Schwarz inequalities it follows

$$\|\Delta \mathbf{e}_j\|_2 \leq \|\mathbf{A}\|_2^{j-\hat{t}_i} \|\Delta \mathbf{e}_{\hat{t}_i}\|_2 + \sum_{l=1}^{j-\hat{t}_i} \|\mathbf{A}\|_2^{j-\hat{t}_i-l} \|\mathbf{K}\|_2. \quad (4.11)$$

Hence, the expected estimation error vector  $\Delta \mathbf{e}_j$  is bounded for any  $j$ , and the system is not perfectly attackable.  $\square$

Following results in the proof of Theorem 9, we can specify even stronger condition on  $f$ , given in the following corollary.

**Corollary 10.** *Consider an LTI system from (2.1) with a global data integrity policy  $(\mu, f, L)$ , where  $L$  is finite and  $f$  is observability index of  $(\mathbf{J}_1, \tilde{\mathbf{C}}_1)$  pair, with  $\mathbf{J}_1$  and  $\tilde{\mathbf{C}}_1$  from (4.6). Then the system is not perfectly attackable.*

*Proof.* Follows from the proofs of Theorems 7 and 9.  $\square$

Theorem 7 assumes that the attacker has the full knowledge of the system's integrity enforcement policy – i.e., at which time-points integrity enforcements will occur. As we illustrate in Chapter 7, this allows the attacker to plan attacks that maximize the error, while ensuring stealthiness of the attack by reducing the state estimation errors to the levels that

will not trigger detection during integrity enforcement intervals. On the other hand, if the attacker does not have the knowledge of  $\mu$  (i.e., if (s)he is not aware of the time points in which integrity enforcements would occur), the integrity enforcement requirements can be additionally relaxed; as the attacker does not know when enforcements occur, (s)he has to ensure that if at any future point (including the next time-step) malicious data cannot be injected, the residue would still remain below the threshold (3.14). Thus, we obtain the following result.

**Theorem 11.** *Any LTI system from (2.1) with a global data integrity policy  $(\mu, 1, L)$  (i.e., with  $f = 1$ ) is not perfectly attackable for any stealthy attacker that does not know the time points  $\mu$  when data integrity is enforced.*

*Proof.* First, note that the sequence  $\mathbf{CA}\Delta\mathbf{e}_k$  cannot be bounded if the attacker wants to introduce an unbounded state estimation error. If it was bounded, from (2.11) and (3.14) it would follow that  $\mathbf{a}_k$  is always bounded; this in turn would imply that the system from (2.10) has bounded inputs, which since matrix  $(\mathbf{A} - \mathbf{KCA})$  is stable ( $\mathbf{K}$  is Kalman gain) would imply that  $\Delta\mathbf{e}_k$  cannot diverge – the reachable set  $\mathcal{R}$  can not be unbounded.

On the other hand, let us assume that the system is perfectly attackable – i.e., the expected state estimation error can be unbounded. Then, from our previous argument it follows that  $\mathbf{CA}\Delta\mathbf{e}_k$  is unbounded and thus we can find  $k$  and  $\Delta\mathbf{e}_k$  such that  $\|\mathbf{CA}\Delta\mathbf{e}_k\|_2 > 1$ . Then, if global data integrity is enforced only once at the time-step  $k + 1$ , from (2.11) it would follow that  $\|\Delta\mathbf{z}_{k+1}\| = \|\mathbf{CA}\Delta\mathbf{e}_k\|_2 > 1$ , which violates the stealthiness requirement from (3.14).  $\square$

Theorems 7 and 9 consider a worst-case scenario without any constraints or assumptions about the set of compromised sensors  $\mathcal{K}$  (e.g., that less than  $q$  sensors are compromised). Yet, some knowledge about the set  $\mathcal{K}$  may be available at design-time. For instance, for *MitM* attacks some sensors cannot be in set  $\mathcal{K}$ , such as on-board sensors

that do not communicate over a network to deliver information to the estimator, or sensors with built-in continuous data authentication. In these cases, the number of integrity enforcements can be reduced.

**Corollary 12.** *Consider a system from (2.1) with a global data integrity policy  $(\mu, f, L)$ , where  $f = \min(\psi, q_{un}^*)$ ,  $\psi$  is the observability index of  $(\mathbf{A}, \mathbf{C})$ , and  $q_{un}^*$  denotes the number of unstable eigenvalues  $\lambda_i$  of  $\mathbf{A}$  for which the corresponding eigenvector  $\mathbf{v}_i$  satisfies  $\text{supp}(\mathbf{C}\mathbf{v}_i) \in \mathcal{K}$ . Then the system is not perfectly attackable.*

*Proof.* The proof directly follows the proof of Theorem 7, with the only difference that all  $\alpha_i \rightarrow \infty$  from Lemma 8 also have to correspond to the unstable eigenvectors  $\mathbf{v}_i$  satisfying that  $\text{supp}(\mathbf{C}\mathbf{v}_i) \in \mathcal{K}$ ; otherwise, consider  $\alpha_i \rightarrow \infty$ , from a decomposition of a ‘large’  $\Delta \mathbf{e}_k$  such that  $\|\mathbf{C}\alpha_i \mathbf{v}_i\|_2 \rightarrow \infty$  and  $\text{supp}(\mathbf{C}\mathbf{v}_i) \notin \mathcal{K}$ . Then the components of the residue  $\Delta \mathbf{z}_{k+1}$  whose indices are in  $\text{supp}(\mathbf{C}\mathbf{v}_i)$  but not in  $\mathcal{K}$  (i.e.,  $P_{\text{supp}(\mathbf{C}\mathbf{v}_i) \setminus \mathcal{K}} \Delta \mathbf{z}_{k+1}$ ) cannot be influenced by the attack signal  $\mathbf{a}_{k+1}$ , meaning that their large values due to  $\alpha_i \rightarrow \infty$  cannot be compensated for by the attack signal, and thus will violate the stealthiness condition (3.14).  $\square$

Let us recall Definition 1 that introduced  $\mathcal{A}_k$ , the set of all stealthy attacks up to time  $k$  – it only requires that attack vector  $\mathbf{a}_{1..k} \in \mathcal{A}_k$  satisfies the stealthiness conditions up to time  $k$ . Thus, as shown in the proof of Theorem 7, the attacker applying attack  $\mathbf{a}_{1..k} \in \mathcal{A}_k$  may have to violate the stealthiness constraint during the next integrity enforcement block, since for those time-points  $t$  when integrity is enforced  $\mathbf{a}_t = \mathbf{0}$ . As the attacker’s goal is to remain stealthy even during integrity enforcements, we consider policy-aware stealthy attack sets.

**Definition 5.** *For an integrity enforcement policy  $(\mu, f, L)$ , the set of all policy-aware stealthy attacks up to time  $k$  is*

$$\mathcal{A}_k^\mu = \left\{ \mathbf{a}_{1..k} \mid \left. \begin{array}{l} \mathbf{a}_{1..k'} \in \mathcal{A}_{k'}, \\ k' = \min \{t \mid (t - f + 1 \in \mu) \wedge (t \geq k)\} \end{array} \right\} \right\}.$$

Intuitively, the attacker will always plan attacks at least until the end of next integrity enforcement block (captured by  $k'$ ), while keeping the probability of detection as low. Thus, we also need to modify the definition of the  $k$ -reachable region  $\mathcal{R}_k$  (Def. 2), as it depends on the employed set of stealthy attacks.

**Definition 6.** *The policy aware  $k$ -reachable region  $\mathcal{R}_k^\mu$  of the state estimation error under the attack (i.e.,  $\mathbf{e}_k^a$ ) is the set*

$$\mathcal{R}_k^\mu = \left\{ \mathbf{e} \in \mathbb{R}^n \mid \begin{array}{l} \mathbf{e}\mathbf{e}^T \preceq E[\mathbf{e}_k^a]E[\mathbf{e}_k^a]^T + \gamma \text{Cov}(\mathbf{e}_k^a), \\ \mathbf{e}_k^a = \mathbf{e}_k^a(\mathbf{a}_{1..k}), \mathbf{a}_{1..k} \in \mathcal{A}_k^\mu \end{array} \right\}. \quad (4.12)$$

Furthermore, the global policy-aware reachable region  $\mathcal{R}^\mu$  of the state estimation error  $\mathbf{e}_k^a$  is the set

$$\mathcal{R}^\mu = \bigcup_{k=0}^{\infty} \mathcal{R}_k^\mu. \quad (4.13)$$

The above definition introduces a region that can be reached by an attacker that both considers past behavior and plans accordingly into the future to avoid being detected. Since  $\mathcal{A}_k^\mu \subseteq \mathcal{A}_k$ , it directly follows that  $\mathcal{R}^\mu \subseteq \mathcal{R}$ , and the boundedness property holds. Finally, note that when no integrity enforcements are used it follows that  $\mathcal{R}^\mu \equiv \mathcal{R}$ .

#### 4.0.1 Guarantees with Sensor-wise Integrity Enforcement

We now consider the case where the system has one unstable eigenvalue  $\lambda_1$  with the corresponding eigenvector  $\mathbf{v}_1$ , but the result can be generalized. Also, let's assume that all sensor integrity enforcement policies use  $f_i = 1$  and have  $t_k^i = t_k^{i+1} - 1$  for all  $k$  and all  $i = 1, \dots, p - 1$  (i.e., sensors enforce integrity in consecutive points, first  $s_1$ , then  $s_2$ , etc); this also implies all  $L_i$  are equal.

It can be shown that the system is not perfectly attackable in this case. The proof follows the ideas from the proofs of Theorems 7 and 9. If  $s_1$  integrity is enforced at  $t_k^1 = j$ , that would mean that  $\Delta \mathbf{z}_{j,1} = P_{\{s_1\}} \mathbf{C} \mathbf{A} \Delta \mathbf{e}_{j-1} = P_{\{s_1\}} \tilde{\mathbf{C}} \mathbf{J} \alpha_{1..n}$  as in Theorem 9,

and thus  $\|P_{\{s_1\}}\tilde{\mathbf{C}}\mathbf{J}\alpha_{1..n}\|_2 \leq 1$ . From Lemma 8, if  $\Delta\mathbf{e}_{j-1}$  is unbounded, only  $\alpha_1 \rightarrow \infty$ , and thus  $\mathbf{J}_1$  as in (4.8) is scalar. To account for this,  $P_{\{s_1\}}\mathbf{C}\mathbf{v}_1\lambda_1\alpha_1$  has to be zero, which implies that  $\mathbf{v}_1 \in \text{null}(P_{\{s_1\}}\mathbf{C})$ . Similarly, it can be shown that from  $\Delta\mathbf{z}_{j,i}$ , it follows that  $\mathbf{v}_1 \in \text{null}(P_{\{s_i\}}\mathbf{C})$  for  $1 \leq i \leq n$ . This can be represented as  $\mathbf{C}\mathbf{v}_1 = \mathbf{0}$ , which since  $\lambda_1 \neq 0$  implies that  $\mathbf{v}_1 \in \text{null}(\mathbf{C})$ . This is a contradiction because  $(\mathbf{A}, \mathbf{C})$  is observable from our initial assumptions.

## Analysis and Design of Safe Integrity Enforcement Policies

In the previous chapter, we have shown that with even intermittent integrity enforcements a stealthy attacker cannot introduce an unbounded state estimation error, irrelevant of the set of compromised sensors  $\mathcal{K}$ . However, we still need to provide methods to evaluate whether a specific integrity enforcement policy ensures the desired estimation performance (i.e., state estimation error) even in the presence of attacks. Furthermore, our goal is to also provide a design framework to derive integrity enforcement policies that ensure that the state estimation errors remain within a desired region even under attack. Thus, in this chapter, we introduce a computationally efficient method to achieve this based on an efficient estimation of the reachable region  $\mathcal{R}_k^\mu$  from (4.12) for systems with intermittent data integrity enforcements.

### 5.1 $\chi^2$ detector

While several metrics can be used to capture safety requirements, we require that the norm of the induced estimation error is always below a pre-specified safe threshold value  $\rho$ . The basis for our analysis is a method (presented in the following subsection) to estimate

reachable region  $\hat{\mathcal{R}}^\alpha$  (as well as  $k$ -reachable regions  $\hat{\mathcal{R}}_k^\alpha$ ); the proposed method allows capturing effects of integrity enforcements (i.e.,  $\mathbf{a}_k = \mathbf{0}$ ) without introducing a high level of overapproximation that would result in more frequent integrity enforcements.

Determining safe integrity enforcement policies  $(\mu, f, L)$  is more challenging. Parameter  $f$  directly follows from (6.18). We determine elements of  $\mu$  in iterative manner. First, we determine  $t_1 \in \mu$  as the maximal time such that the first integrity enforcement at  $t_1$  causes the attacker to reduce estimation error before  $\|\Delta \mathbf{e}_k\|_2$  reaches  $\rho$ .<sup>1</sup> We similarly obtain  $t_2 \in \mu$ , but starting from  $\mathcal{R}_{t_1}$  as the initial region. Note that we do not search through all possible  $t_2 - t_1$ ; we rather evaluate candidates obtained as the minimal time an overapproximation similar to (4.11) needs to reach the safety threshold and return to the initial region. When this method was repeated, we observed that the time between starts of consecutive blocks would quickly settle and the policy would effectively move to periodic enforcement blocks.

Two caveats are in order. First, while this procedure is computationally heavy, it is performed at design-time (as opposed to runtime). Also, while the proposed policy would ensure system safety even in the presence of attack, providing optimal policies (e.g., that minimize average frequency of integrity enforcements) is an avenue for future work.

### 5.1.1 Evaluation of the State Estimation Error Regions

There exist algorithms that approximate estimation error regions (e.g., [Mo and Sinopoli (2016)]), but with significant limitations due to the level of overapproximation or the fact that they do not support analysis with integrity enforcement in addition to the stealthiness constraints. Thus, we developed a method to estimate reachable regions in our case.

As  $\|\Delta \mathbf{z}_k\|_\infty \leq \|\Delta \mathbf{z}_k\|_2 \leq \|\Delta \mathbf{z}_k\|_{\mathbf{P}^{-1}} |\lambda_{max}| \leq \alpha |\lambda_{max}|$ , the  $k$ -reachable regions can be overapproximated by capturing the stealthiness constraint as  $\|\Delta \mathbf{z}_k\|_\infty \leq \alpha |\lambda_{max}|$ . Due to

---

<sup>1</sup> The attacker may breach the threshold but would not be able to remain stealthy. To also ensure that the stealthiness constraint is violated before the bound is breached, a lower threshold should be used to generate the policy.

linearity of the constraints, we set the constraint to be  $\|\Delta \mathbf{z}_k\|_\infty \leq 1$ , and multiply obtained values by  $\alpha|\lambda_{max}|$  after. Thus, the system and attacker have to satisfy

$$\begin{aligned} \Delta \mathbf{e}_k &= - \underbrace{\left[ (\mathbf{A} - \mathbf{KCA})^{k-1} \mathbf{K} \mid \dots \mid \mathbf{K} \right]}_{\mathbf{M}_k} \mathbf{a}_{1..k} \\ \Delta \mathbf{z}_k &= \underbrace{\left[ -\mathbf{CAM}_{k-1} \mid \mathbf{I} \right]}_{\mathbf{N}_k} \mathbf{a}_{1..k} \\ |\Delta \mathbf{z}_{j,i}| &\leq 1, \quad i \in \{1, \dots, p\}, j \in \{1, \dots, k\}, \end{aligned} \quad (5.1)$$

where  $\mathbf{a}_{1..k} = [(\mathbf{a}_1)^T \dots (\mathbf{a}_k)^T]^T$ ; this can be summarized as

$$\underbrace{\left[ \begin{array}{c|c} \mathbf{I}_{n+pk} & \begin{array}{c} \hline \mathbf{M}_k P_Q^\dagger \\ -\mathbf{N}_1 P_{\mathcal{K}}^\dagger \mid \mathbf{0}_{p \times (k-1)q} \\ \hline \dots \\ \hline -\mathbf{N}_k P_Q^\dagger \\ \hline \end{array} \\ \hline \mathbf{0}_{2pk \times n} & \begin{array}{c} \mathbf{I}_{kp} \\ -\mathbf{I}_{kp} \end{array} \mid \mathbf{0}_{2pk \times kq} \end{array} \right]}_{(\Omega_k)_{(n+3pk) \times (n+pk+kq)}} \underbrace{\begin{bmatrix} \Delta \mathbf{e}_k \\ \Delta \mathbf{z}_1 \\ \dots \\ \Delta \mathbf{z}_k \\ P_{\mathcal{K}} \mathbf{a}_1 \\ \dots \\ P_{\mathcal{K}} \mathbf{a}_k \end{bmatrix}}_{\mathbf{r}_k^{rez}} \geq \underbrace{\begin{bmatrix} \mathbf{0}_{n+pk} \\ -\mathbf{1}_{2pk} \end{bmatrix}}_{\mathbf{b}_k} \quad (5.2)$$

Here,  $P_{\mathcal{K}} \in \mathbb{R}^{|\mathcal{K}| \times p}$  is the projection matrix that keeps only elements from the set  $\mathcal{K}$  ( $\mathcal{K} \subseteq S$ ), and  $P_Q$  is block-diagonal with  $k$  matrices  $P_{\mathcal{K}}$  on the diagonal.

Let us introduce a  $k$ -reachable region  $\hat{\mathcal{R}}_k^\alpha$  as in (3.2) with one difference - instead of the  $\|\Delta \mathbf{z}_k\|_{p-1} \leq \alpha$  requirement, we impose that  $\|\Delta \mathbf{z}_k\|_\infty \leq \alpha|\lambda_{max}|$ . In addition, we introduce  $\hat{\mathcal{R}}^\alpha = \cup_{k=0}^\infty \hat{\mathcal{R}}_k^\alpha$ . Since,  $\|\Delta \mathbf{z}_k\|_{p-1} \leq \alpha \Rightarrow \|\Delta \mathbf{z}_k\|_\infty \leq \alpha|\lambda_{max}|$  it follows that  $\hat{\mathcal{R}}^\alpha \subseteq \hat{\mathcal{R}}^\alpha$ , and we use  $\hat{\mathcal{R}}^\alpha$  to bound  $\hat{\mathcal{R}}^\alpha$ .

From (5.2),  $\hat{\mathcal{R}}_k^\alpha$  is a polyhedron in  $\mathbb{R}^n$ . Note that the maximal value of  $\|\Delta \mathbf{e}_k\|_2$  over a polyhedron can be obtained in a vertex of the polyhedron [Boyd and Vandenberghe (2004)]. The vertices of  $\hat{\mathcal{R}}_k^\alpha$  satisfy that  $kq$  constraints from (5.5) are active. This means that all equalities and  $kq$  inequalities from (5.5) are active rows in (5.2). We define matrix  $\Omega_k^{act}$  that contains all active rows of  $\Omega_k$  from (5.2). Let  $\{(\Omega_k^{act})^1, \dots, (\Omega_k^{act})^{full}\}$  be

the set of all such  $\Omega_k^{act}$  with the full rank. Then, if  $(\mathbf{b}_k^{act})^i, 1 \leq i \leq full$  represent the corresponding values from  $\mathbf{b}_k$ , we define

$$(\mathbf{r}_k^{rez})^i = ((\Omega_k^{act})^i)^\dagger (\mathbf{b}_k^{act})^i \quad (5.3)$$

where  $((\Omega_k^{act})^i)^\dagger$  denotes the pseudoinverse matrix of  $(\Omega_k^{act})^i$ . Thus, the set of vertices of  $\hat{\mathcal{R}}_k^\alpha$  can be expressed as  $\{(\mathbf{r}_k^{rez})^i : (i \in \{1, 2, \dots, full\}) \wedge ((\mathbf{r}_k^{rez})^i \text{ satisfies (5.2)})\}$ .

Finally, to determine the vertices of  $\hat{\mathcal{R}}_k^\alpha$  in case when integrity is enforced at time points  $H \subseteq \{1, \dots, k\}$ , we add  $\mathbf{a}_j = \mathbf{0}, \forall j \in H$  to the system (5.5), and repeat the process.

The above procedure provides an estimate of the maximal estimation error in each step  $k$ . On one hand, the computation time grows exponentially with  $k$  and calculations for higher numbers of steps could become unfeasible. On the other hand, since we evaluate reachable state estimation errors to provide guidance on the effects and design of integrity enforcement policies, we did not face limitations caused by the computation times for analyzed systems; even after first integrity enforcements we would observe that new  $\hat{\mathcal{R}}_k^\alpha \subset \cup_{i=0}^{k-1} \hat{\mathcal{R}}_i^\alpha$ , which was exploited to reduce problem size.

In addition to the case study presented in Chapter 7, we evaluated the proposed reachable region estimation method on a simple vehicle model from [Mo and Sinopoli (2016)], and analyzed the level of over-approximation due to the use of  $\infty$ -norm. For example, as shown in Figure 5.1, the 4-step reachable region obtained by our method is a very good approximation of the actual reachability region (which was obtained with the use of 2-norm and brute-force discretization of the state space). Similar results were obtained for other reachable regions.

We also compared our method to the algorithm introduced in [Mo and Sinopoli (2016)] that recursively over- and under-approximates the estimation error with outer and inner ellipsoids. Although the method from [Mo and Sinopoli (2016)] requires lower computation time, it does not directly allow for capturing the effects of integrity enforcement, since  $\mathbf{a}_k$  plays no part in the procedure. It may also make arbitrarily large over-approximation for

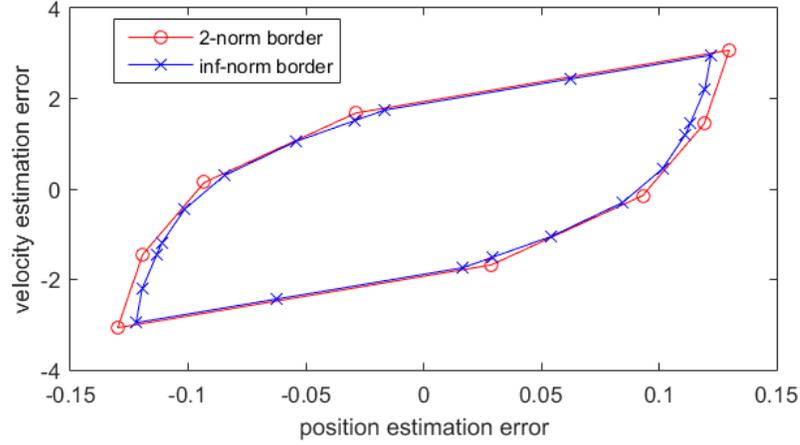


FIGURE 5.1: 4-step reachable regions for 2-norm and  $\infty$ -norm.

the outer ellipsoids (which we are interested in due to the safety requirements), depending on the shape of the actual region. For example, from the 4-step outer ellipsoidal region estimations computed in [Mo and Sinopoli (2016)], the bounds on the maximal reachable state estimation errors are approximately 1.5 times for position and 3 times for velocity more conservative than our estimations. This increase in the reachable region overapproximation would double the frequency of integrity enforcement. Still, if conservativeness of an existing reachability analysis (e.g., from [Mo and Sinopoli (2016)]) is not a concern it could be used if extended to allow integrity enforcements ( $\mathbf{a}_k = \mathbf{0}$ ).

## 5.2 Reachable State Estimation Errors with Intermittent Integrity Enforcements (CUSUM)

Consider an LTI system from (2.1), (2.9) with a global data integrity policy  $(\mu, f, L)$ . As in Definition 5, we use  $\mathbf{a}_{1..k} = [(\mathbf{a}_1)^T \dots (\mathbf{a}_k)^T]^T \in \mathbb{R}^{pk}$  to capture attack vectors up to step  $k$ , where  $\text{supp}(\mathbf{a}_j) = \tilde{\mathcal{K}}_j$ ,  $j = 1, \dots, k$ , and

$$\tilde{\mathcal{K}}_j = \begin{cases} \emptyset, & j - i \in \mu, \text{ for some } i, 0 \leq i < f, \\ \mathcal{K}, & \text{otherwise} \end{cases}$$

Here,  $\tilde{\mathcal{K}}_j$  captures the set of compromised sensor measurements received in step  $j$  – i.e., if data integrity is enforced at step  $j$  then no measurements are compromised. In addition, let us define  $\text{supp}(\mathbf{a}_{1..k}) = \mathcal{Q}_k \subseteq \{1, \dots, pk\}$ ; note that  $\mathcal{Q}_k$  effectively captures information about the applied integrity enforcement policy, and

$$|\mathcal{Q}_k| = |\mathcal{Q}_{k-1}| + |\tilde{\mathcal{K}}_k| = \sum_{i=1}^k |\tilde{\mathcal{K}}_i|. \quad (5.4)$$

From (2.10) and (2.11),  $\Delta \mathbf{e}_k$  and  $\Delta \mathbf{z}_k$  can be captured in a non-recursive form as

$$\begin{aligned} \Delta \mathbf{e}_k &= - \underbrace{\left[ (\mathbf{A} - \mathbf{KCA})^{k-1} \mathbf{K} \mid \dots \mid \mathbf{K} \right]}_{\mathbf{M}_k} \mathbf{a}_{1..k} \\ \Delta \mathbf{z}_k &= \underbrace{\left[ -\mathbf{CAM}_{k-1} \mid \mathbf{I} \right]}_{\mathbf{N}_k} \mathbf{a}_{1..k} \end{aligned} \quad (5.5)$$

To incorporate the information about the sparsity of the attack vector, we use suitable projections onto  $\mathcal{Q}_k$  and  $\tilde{\mathcal{K}}_1, \dots, \tilde{\mathcal{K}}_k$ , which satisfy  $\mathbf{P}_{\mathcal{Q}_k} = \text{BlckDiag}(\mathbf{P}_{\tilde{\mathcal{K}}_1}, \dots, \mathbf{P}_{\tilde{\mathcal{K}}_k})$ . In addition, it holds that  $\mathbf{P}_{\tilde{\mathcal{K}}_j}^\dagger = \mathbf{P}_{\tilde{\mathcal{K}}_j}^T$ , since  $\mathbf{P}_{\tilde{\mathcal{K}}_j} \mathbf{P}_{\tilde{\mathcal{K}}_j}^T = \mathbf{I}_{|\tilde{\mathcal{K}}_j|}$ , for  $j = 1, \dots, k$ , and thus  $\mathbf{P}_{\mathcal{Q}_k}^\dagger = \mathbf{P}_{\mathcal{Q}_k}^T$ . Then, (5.5) can be restated as

$$\begin{aligned} \Delta \mathbf{e}_k &= - \underbrace{\left[ (\mathbf{A} - \mathbf{KCA})^{k-1} \mathbf{K} \mathbf{P}_{\tilde{\mathcal{K}}_1}^\dagger \mid \dots \mid \mathbf{K} \mathbf{P}_{\tilde{\mathcal{K}}_k}^\dagger \right]}_{\mathbf{M}_k \mathbf{P}_{\mathcal{Q}_k}^\dagger} \mathbf{P}_{\mathcal{Q}_k} \mathbf{a}_{1..k} \\ \Delta \mathbf{z}_k &= \underbrace{\left[ -\mathbf{CAM}_{k-1} \mathbf{P}_{\mathcal{Q}_{k-1}}^\dagger \mid \mathbf{P}_{\tilde{\mathcal{K}}_k}^\dagger \right]}_{\mathbf{N}_k \mathbf{P}_{\mathcal{Q}_k}^\dagger} \mathbf{P}_{\mathcal{Q}_k} \mathbf{a}_{1..k} \end{aligned} \quad (5.6)$$

with matrices  $\mathbf{M}_k \mathbf{P}_{\mathcal{Q}_k}^\dagger$  and  $\mathbf{N}_k \mathbf{P}_{\mathcal{Q}_k}^\dagger$  capturing information about the time steps in which data integrity is enforced.

For the general form of the detection function  $g_k$  it may not be possible to obtain a simple analytical solution for the regions  $\mathcal{R}_k^\mu$  and  $\mathcal{R}^\mu$ . Therefore, in this chapter we will

focus on a specific detection function employed by cumulative sum (CUSUM) [Granjon (2014)] detectors. However, the presented method can be extended in similar fashion to cover other detectors, such as SPRT [Kwon et al. (2015)] and generalized likelihood test. CUSUM observes two hypothesis,  $\mathcal{H}_0 : \mathbf{z}_k \sim \mathcal{N}(0, \mathbf{Q})$  and  $\mathcal{H}_1 : \mathbf{z}_k \sim \mathcal{N}(\mathbf{a}_k, \mathbf{Q})$ . One issue that arises from using CUSUM is its non-linearity, given that it accumulates the data until decision is reached, after which observation window is reset. In addition, the exact distribution for  $\mathbf{z}_k$  under  $\mathcal{H}_1$  is not known since the mean of compromised  $\mathbf{z}_k$  (i.e.,  $\mathbf{a}_k$ ) changes over time, which causes suboptimal performance of CUSUM detector [Granjon (2014)]. To address the first issue, let us consider the stealthiness condition of the attacker (i.e. no alarms). Due to this condition the attacker will avoid crossing detection threshold  $h$ , determined such that false positive rate is kept sufficiently low as to avoid frequent system recovery procedures, and thus avoiding non-linear resets of CUSUM. In essence, if the CUSUM resets, the attacker is detected, and stealthiness condition is invalidated.

The second challenge, i.e. unknown distribution of  $\mathcal{H}_1$  is fairly common in practical applications, but CUSUM is still used as it provides satisfactory performance [Granjon (2014)]. We approximate the detection function by initializing log-likelihood ratio  $\Lambda_k \equiv 0$  when the system is not under the attack, as previously proposed for SPRT detectors in e.g., [Kwon et al. (2015); Kwon and Hwang (2017)]; this will ensure that  $g_k$  does not go above the threshold without attack. Consequently, from these assumptions, it follows that the detector function of CUSUM detector can be captured as

$$\begin{aligned}
g_k &= g_{k-1} + \Lambda_k = \sum_{\tau=1}^k \left( \frac{1}{2} \mathbf{z}_\tau^T \mathbf{Q}^{-1} \mathbf{z}_\tau + \log c \sqrt{(2\pi)^p \det(\mathbf{Q})} \right) = \\
&= \frac{1}{2} \sum_{\tau=1}^k (\mathbf{z}_\tau^T \mathbf{Q}^{-1} \mathbf{z}_\tau) + k \log c \sqrt{(2\pi)^p \det(\mathbf{Q})}
\end{aligned} \tag{5.7}$$

where  $\Lambda_k = \log \frac{f_a(\mathbf{z}_k)}{f(\mathbf{z}_k)}$ ,  $f_a$  and  $f$  are probability density functions of the residuals under the attack and in regular operation respectively, and  $c = e^{-\frac{p}{2}} / \sqrt{(2\pi)^p \det(\mathbf{Q})}$  is a design

constant initialized such that log-likelihood ratio  $\Lambda_k \equiv 0$ . Thus, in this case the attacker's stealthiness constraint from (2.16) (i.e.,  $P(g_k^a > h) \leq P(g_k > h) + \varepsilon$ ) can be captured as

$$P\left(\sum_{\tau=1}^k (\mathbf{z}_\tau^{aT} \mathbf{Q}^{-1} \mathbf{z}_\tau^a) > 2h + kp\right) \leq \varepsilon + P\left(\sum_{\tau=1}^k (\mathbf{z}_\tau^T \mathbf{Q}^{-1} \mathbf{z}_\tau) > 2h + kp\right)$$

Given that these two sums have the non-central  $\chi^2$  (left) and (central)  $\chi^2$  distributions, from Theorem 1 and the proof of Lemma 3 it follows that the above constraint is equivalent to

$$\sqrt{\sum_{\tau=1}^k \|\Delta \mathbf{z}_\tau\|_{\mathbf{Q}^{-1}}^2} \leq \alpha_{\chi^2}(\varepsilon, kp, 2h + kp). \quad (5.8)$$

On the other hand, from (5.6) it follows that

$$\begin{aligned} \sum_{\tau=1}^k \|\Delta \mathbf{z}_\tau\|_{\mathbf{Q}^{-1}}^2 &= \sum_{\tau=1}^k \Delta \mathbf{z}_\tau^T \mathbf{Q}^{-1} \Delta \mathbf{z}_\tau = \\ &= \sum_{\tau=1}^k (\mathbf{P}_{\mathcal{Q}_k} \mathbf{a}_{1..k})^T [\mathbf{N}_\tau \mathbf{P}_{\mathcal{Q}_\tau^\dagger} \mathbf{0}_{p \times (|\mathcal{Q}_k| - |\mathcal{Q}_\tau|)}]^T \mathbf{Q}^{-1} \\ &\quad [\mathbf{N}_\tau \mathbf{P}_{\mathcal{Q}_\tau^\dagger} \mathbf{0}_{p \times (|\mathcal{Q}_k| - |\mathcal{Q}_\tau|)}] \mathbf{P}_{\mathcal{Q}_k} \mathbf{a}_{1..k}. \end{aligned}$$

Hence, from (5.8), the attacker's stealthiness constraint under considered integrity enforcement policy  $\mu$  can be captured as

$$\|P_{\mathcal{Q}_k} \mathbf{a}_{1..k}\|_{\Theta_k} \leq \alpha_{\chi^2}(\varepsilon, kp, 2h + kp), \quad (5.9)$$

where

$$\Theta_k = \sum_{\tau=1}^k [\mathbf{N}_\tau \mathbf{P}_{\mathcal{Q}_\tau^\dagger} \mathbf{0}_{p \times (|\mathcal{Q}_k| - |\mathcal{Q}_\tau|)}]^T \mathbf{Q}^{-1} [\mathbf{N}_\tau \mathbf{P}_{\mathcal{Q}_\tau^\dagger} \mathbf{0}_{p \times (|\mathcal{Q}_k| - |\mathcal{Q}_\tau|)}]. \quad (5.10)$$

For the above matrix  $\Theta_k$ , the following property holds.

**Lemma 13.** *For any  $k \geq 1$ , the matrix  $\Theta_k$  is positive definite.*

$$\tilde{\Theta}_k = \begin{bmatrix} (\mathbf{C}\mathbf{A}\mathbf{M}_{k-1}\mathbf{P}_{\mathcal{Q}_{k-1}}^\dagger)^T \mathbf{Q}^{-1} \mathbf{C}\mathbf{A}\mathbf{M}_{k-1}\mathbf{P}_{\mathcal{Q}_{k-1}}^\dagger & (\mathbf{C}\mathbf{A}\mathbf{M}_{k-1}\mathbf{P}_{\mathcal{Q}_{k-1}}^\dagger)^T \mathbf{Q}^{-1} \mathbf{P}_{\tilde{\mathcal{K}}_k}^\dagger \\ (\mathbf{P}_{\tilde{\mathcal{K}}_k}^\dagger)^T \mathbf{Q}^{-1} \mathbf{C}\mathbf{A}\mathbf{M}_{k-1}\mathbf{P}_{\mathcal{Q}_{k-1}}^\dagger & (\mathbf{P}_{\tilde{\mathcal{K}}_k}^\dagger)^T \mathbf{Q}^{-1} \mathbf{P}_{\tilde{\mathcal{K}}_k}^\dagger \end{bmatrix} \quad (5.12)$$

*Proof.* We start with the case when  $k = 1$ . From Definition 4, data integrity is not enforced at  $k = 1$  and thus  $\mathcal{Q}_1 = \tilde{\mathcal{K}}_1 = \mathcal{K}$ . Due to the way projection matrices are formed, we have that

$$\mathbf{P}_{\mathcal{Q}_1}^{\dagger T} \mathbf{P}_{\mathcal{Q}_1}^\dagger = \mathbf{I}_{|\mathcal{Q}_1|} > 0 \quad \text{and} \quad \Theta_1 = [\mathbf{P}_{\mathcal{Q}_1}^\dagger]^T \mathbf{Q}^{-1} [\mathbf{P}_{\mathcal{Q}_1}^\dagger].$$

Since  $\mathbf{Q} > 0$ , it follows that  $\Theta_1 > 0$  as well.

Now, consider the case  $k \geq 2$  and let us assume that  $\Theta_{k-1}$  is positive definite. From (5.10) it follows that

$$\Theta_k = \underbrace{\begin{bmatrix} \Theta_{k-1} & \mathbf{0}_{|\mathcal{Q}_{k-1}| \times |\tilde{\mathcal{K}}_k|} \\ \mathbf{0}_{|\tilde{\mathcal{K}}_k| \times |\mathcal{Q}_{k-1}|} & \mathbf{0}_{|\tilde{\mathcal{K}}_k| \times |\tilde{\mathcal{K}}_k|} \end{bmatrix}}_{\tilde{\Theta}_{k-1}} + \underbrace{[\mathbf{N}_k \mathbf{P}_{\mathcal{Q}_k}^\dagger]^T \mathbf{Q}^{-1} [\mathbf{N}_k \mathbf{P}_{\mathcal{Q}_k}^\dagger]}_{\tilde{\Theta}_k} \quad (5.11)$$

and we consider the following two cases.

Case I: There does not exist  $i$ , such that  $0 \leq i < f$  and  $k - i \in \mu$ ; this implies that integrity is not enforced at the step  $k$  and  $\tilde{\mathcal{K}}_k = \mathcal{K}$ . Because both  $\Theta_{k-1} > 0$  and  $\mathbf{Q} > 0$ , both addends in (5.11) are positive semidefinite matrices, and  $\Theta_k \geq 0$ . In addition, since  $\Theta_{k-1}$  is positive definite by assumption,  $\text{null}(\tilde{\Theta}_{k-1}) = \mathcal{R}([\mathbf{0}_{|\tilde{\mathcal{K}}_k| \times |\mathcal{Q}_{k-1}|} \quad \mathbf{I}_{|\tilde{\mathcal{K}}_k|}]^T)$ . Furthermore, from (5.6), we have (5.12). Given that  $(\mathbf{P}_{\tilde{\mathcal{K}}_k}^\dagger)^T \mathbf{Q}^{-1} \mathbf{P}_{\tilde{\mathcal{K}}_k}^\dagger > 0$ , it follows that  $\text{null}(\tilde{\Theta}_k)$  cannot have non-zero vectors from  $\mathcal{R}([\mathbf{0}_{|\tilde{\mathcal{K}}_k| \times |\mathcal{Q}_{k-1}|} \quad \mathbf{I}_{|\tilde{\mathcal{K}}_k|}]^T)$ . Therefore,

$$\text{null}(\tilde{\Theta}_k) \cap \text{null}(\tilde{\Theta}_{k-1}) = \{\mathbf{0}\}. \quad (5.13)$$

Now, assume that there exists a non-zero vector  $\mathbf{v}$  such that  $\mathbf{v} \in \text{null}(\Theta_k)$  – i.e.,  $(\tilde{\Theta}_k + \tilde{\Theta}_{k-1})\mathbf{v} = 0$ , and thus

$$\mathbf{v}^T \tilde{\Theta}_k \mathbf{v} = -\mathbf{v}^T \tilde{\Theta}_{k-1} \mathbf{v}.$$

However, since  $\mathbf{v}$  cannot be in the null-spaces of both matrices due to (5.13), and  $\tilde{\Theta}_{k-1}$  and  $\tilde{\Theta}_k$  are both positive semidefinite, this is a clear contradiction. Consequently,  $\Theta_k = \tilde{\Theta}_k + \tilde{\Theta}_{k-1} \text{ null}(\Theta_k) = \{0\}$ , and since  $\Theta_k$  is a positive semidefinite matrix it holds that  $\Theta_k > 0$ .

Case II: There exists  $i$ , such that  $0 \leq i < f$  and  $k - i \in \mu$ ; i.e., integrity is enforced at the step  $k$ . Thus,  $|\tilde{\mathcal{K}}_k| = 0$ , so  $\tilde{\Theta}_{k-1} = \Theta_{k-1}$  is positive definite. Thus, since  $\tilde{\Theta}_k \geq 0$ , it follows that  $\Theta_k = \tilde{\Theta}_k + \tilde{\Theta}_{k-1}$  is positive definite.  $\square$

Now, the specification of the stealthiness condition from (5.9) allows us to obtain the following result.

**Theorem 14.** *The  $k$ -reachable region  $\mathcal{R}_k^\mu$  under a global data integrity enforcement policy  $(\mu, f, L)$  can be represented as*

$$\mathcal{R}_k^\mu = \left\{ \mathbf{e}_k^a | \mathbf{e}_k^a \mathbf{e}_k^{aT} \preceq \alpha_{\chi^2}^2 [\mathbf{M}_k P_{\mathcal{Q}_k}^\dagger \mathbf{0}] \Theta_t^{-1} [\mathbf{M}_k P_{\mathcal{Q}_k}^\dagger \mathbf{0}]^T + \gamma \Sigma \right\} \quad (5.14)$$

where  $\alpha_{\chi^2}^2 = \alpha_{\chi^2}^2(\varepsilon, tp, 2h + tp)$ ,  $t$  is the first end of an integrity enforcement block following  $k$  - i.e., the earliest time point such that  $t - f + 1 \in \mu$  and  $k \leq t$ , and  $\mathbf{0} = \mathbf{0}_{|\mathcal{Q}_k| \times (|\mathcal{Q}_t| - |\mathcal{Q}_k|)}$ .

*Proof.* Consider the stealthiness constraints (5.9) at time  $t$ , which can be written as

$$\alpha_{\chi^2}^2(\varepsilon, tp, 2h + tp) - (\mathbf{P}_{\mathcal{Q}_t} \mathbf{a}_{1..t})^T \Theta_t \mathbf{P}_{\mathcal{Q}_t} \mathbf{a}_{1..t} \geq 0. \quad (5.15)$$

Now, using Schur complement and Lemma 13, we obtain

$$\begin{bmatrix} \Theta_t^{-1} & \mathbf{P}_{\mathcal{Q}_t} \mathbf{a}_{1..t} \\ (\mathbf{P}_{\mathcal{Q}_t} \mathbf{a}_{1..t})^T & \alpha_{\chi^2}^2(\varepsilon, tp, 2h + tp) \end{bmatrix} \succeq 0 \quad (5.16)$$

As the left hand side of (6.11) is positive semidefinite, when multiplied by a matrix from the left, and its transpose from the right, this product will also be positive semidefinite.

If we use the projection matrix  $\mathbf{P}_{\{1, \dots, k, t+1\}}$  for this, we effectively reduce the matrix

$$\begin{aligned}
& \begin{bmatrix} -\mathbf{M}_k \mathbf{P}_{\mathcal{Q}_k}^\dagger & \mathbf{0}_{n \times 1} \\ \mathbf{0}_{1 \times |\mathcal{Q}_k|} & 1 \end{bmatrix} \begin{bmatrix} [\mathbf{I} \ \mathbf{0}] \boldsymbol{\Theta}_t^{-1} [\mathbf{I} \ \mathbf{0}]^T & \mathbf{P}_{\mathcal{Q}_k} \mathbf{a}_{1..k} \\ (\mathbf{P}_{\mathcal{Q}_k} \mathbf{a}_{1..k})^T & \alpha_{\chi^2}^2(\varepsilon, tp, 2h + tp) \end{bmatrix} \begin{bmatrix} -\mathbf{M}_k \mathbf{P}_{\mathcal{Q}_k}^\dagger & \mathbf{0}_{n \times 1} \\ \mathbf{0}_{1 \times |\mathcal{Q}_k|} & 1 \end{bmatrix}^T \geq 0 \iff \\
& \iff \begin{bmatrix} \mathbf{M}_k \mathbf{P}_{\mathcal{Q}_k}^\dagger [\mathbf{I} \ \mathbf{0}] \boldsymbol{\Theta}_t^{-1} [\mathbf{I} \ \mathbf{0}]^T (\mathbf{M}_k \mathbf{P}_{\mathcal{Q}_k}^\dagger)^T & -\mathbf{M}_k \mathbf{P}_{\mathcal{Q}_k}^\dagger \mathbf{P}_{\mathcal{Q}_k} \mathbf{a}_{1..k} \\ -(\mathbf{P}_{\mathcal{Q}_k} \mathbf{a}_{1..k})^T (\mathbf{M}_k \mathbf{P}_{\mathcal{Q}_k}^\dagger)^T & \alpha_{\chi^2}^2(\varepsilon, tp, 2h + tp) \end{bmatrix} \geq 0.
\end{aligned} \tag{5.18}$$

from (6.11) by removing pairs of rows and columns corresponding to  $\mathbf{a}_{k+1..t}$ . Thus, we obtain that

$$\begin{bmatrix} [\mathbf{I}_{|\mathcal{Q}_k|} \ \mathbf{0}] \boldsymbol{\Theta}_t^{-1} [\mathbf{I}_{|\mathcal{Q}_k|} \ \mathbf{0}]^T & \mathbf{P}_{\mathcal{Q}_k} \mathbf{a}_{1..k} \\ (\mathbf{P}_{\mathcal{Q}_k} \mathbf{a}_{1..k})^T & \alpha_{\chi^2}^2(\varepsilon, tp, 2h + tp) \end{bmatrix} \geq 0 \tag{5.17}$$

where  $\mathbf{0} = \mathbf{0}_{|\mathcal{Q}_k| \times (|\mathcal{Q}_t| - |\mathcal{Q}_k|)}$ . Furthermore, with condition (5.17) we need to compute only single  $\boldsymbol{\Theta}_t^{-1}$  for all points between integrity enforcement blocks, as constraints for prior attacks (i.e., time points before  $t$ ) directly follow from (5.17).

The LMI in (6.14) follows from (5.17) as it forms a quadratic representation. We use this specific matrix as it allow us to argue about the stealthiness condition using  $\Delta \mathbf{e}_k$  rather than  $\mathbf{a}_{1..k}$ . Using (5.6) and Schur complement once again, we have

$$[\mathbf{M}_k \mathbf{P}_{\mathcal{Q}_k}^\dagger \ \mathbf{0}] \boldsymbol{\Theta}_t^{-1} [(\mathbf{M}_k \mathbf{P}_{\mathcal{Q}_k}^\dagger \ \mathbf{0})]^T - \frac{1}{\alpha_{\chi^2}^2} \Delta \mathbf{e}_k \Delta \mathbf{e}_k^T \geq 0, \tag{5.19}$$

where  $\alpha_{\chi^2}^2 = \alpha_{\chi^2}^2(\varepsilon, tp, 2h + tp)$ . Hence, from (6.15) and the definition of  $\mathcal{R}_k^\mu$  from (4.12), as well as (2.12) and the fact that  $\text{Cov}[\mathbf{e}_k^a] = \boldsymbol{\Sigma}$ , we finally obtain that (6.4) holds.  $\square$

The representation of the reachable set from (6.4) can be simplified further. Let's define  $\mathbf{Y}_k$  as

$$\mathbf{Y}_k = \alpha_{\chi^2}^2(\varepsilon, tp, 2h + tp) [\mathbf{M}_k \mathbf{P}_{\mathcal{Q}_k}^\dagger \ \mathbf{0}] \boldsymbol{\Theta}_t^{-1} [\mathbf{M}_k \mathbf{P}_{\mathcal{Q}_k}^\dagger \ \mathbf{0}]^T + \gamma \boldsymbol{\Sigma}. \tag{5.20}$$

Then (6.4) is equivalent to  $\mathbf{Y}_k - \mathbf{e}_k^a \mathbf{e}_k^{aT} \geq 0$ , and thus by using Schur complement we

obtain an alternative representation of the  $k$ -reachable regions as

$$\mathcal{R}_k^\mu = \{\mathbf{e}_k^a | \mathbf{e}_k^{aT} \mathbf{Y}_k^{-1} \mathbf{e}_k^a \leq 1\}. \quad (5.21)$$

for the positive definite matrix  $\mathbf{Y}_k$  defined in (5.20). The above representation can be exploited for efficient computation of the reachable-regions.

Furthermore, as we described in Section 2.2, the attacker's goal is to maximize the expected state estimation error  $E[\mathbf{e}_k^a] = \Delta \mathbf{e}_k$ . From the above discussion, the following corollary directly holds by considering the case when  $\gamma = 0$ .

**Corollary 15.** *At any time  $k$ , the maximal norm of the expected state estimation error  $\mathbf{e}_k^a$  caused by the attack satisfies*

$$\max \|E[\mathbf{e}_k^a]\|_2 = \frac{1}{\sqrt{\lambda_{max}(\tilde{\mathbf{Y}}_k)}}, \quad (5.22)$$

where  $\lambda_{max}(\tilde{\mathbf{Y}}_k)$  denotes the largest eigenvalue of the matrix

$$\tilde{\mathbf{Y}}_k = \alpha_{\chi^2}^2(\varepsilon, tp, 2h + tp) [\mathbf{M}_k P_{\mathcal{Q}_k}^\dagger \quad \mathbf{0}] \Theta_t^{-1} [\mathbf{M}_k P_{\mathcal{Q}_k}^\dagger \quad \mathbf{0}]^T,$$

and  $t$  is the next end of integrity enforcement block – i.e., the earliest time point such that  $t - f + 1 \in \mu$  and  $k \leq t$ .

The above corollary provides a very efficient way to evaluate worst-case effects of attacks when an intermittent data integrity enforcement policy is used. By quantifying degradation of the expected state estimation error in the presence of attacks we can analyze the impact of the integrity enforcement policy on limiting the attacker, which can then be used for design of suitable integrity enforcement policies.

### 5.3 Design of Periodic Integrity Enforcement Policies (CUSUM)

For policy design, it is necessary to be able to evaluate impact of an integrity enforcement policy  $\mu$ , not only on reachable regions  $\mathcal{R}_k^\mu$ , for any  $k$ , but even more importantly on  $\mathcal{R}^\mu$

from (4.13). To achieve this, we have to obtain the terminating value  $t$  from Theorem 16, or equivalently from (5.21), such that the reachability analysis can be completed after  $\mathcal{R}_t^\mu$  is obtained – i.e., for which  $\mathcal{R}^\mu = \mathcal{R}_{1..t}^\mu$ , where  $\mathcal{R}_{1..t}^\mu = \bigcup_{k=1}^t \mathcal{R}_k^\mu$ . In the general case, the analysis may never terminate, depending on the particular policy  $(\mu, f, L)$ . Therefore, to simplify the analysis, in this chapter we focus on periodic integrity enforcement policies introduced in Remark 4.

For a periodic integrity enforcement policy  $(\mu, f, L)$ , consider  $t_1$  and  $t_2 = t_1 + L$  time points at which consecutive integrity enforcement blocks end – i.e.,  $t_1 - f + 1 \in \mu$  and  $t_2 - f + 1 \in \mu$ . From the proof of Theorem 16, if the stealthiness requirements from the condition in (6.11) are satisfied at any time  $t \in \mu$ , then they are satisfied for all  $k < t$ , since (5.17) follows from (6.11). Given that  $\mathbf{a}_{t_1-f+1} = \dots = \mathbf{a}_{t_1} = \mathbf{0}$  and  $\mathbf{a}_{t_2-f+1} = \dots = \mathbf{a}_{t_2} = \mathbf{0}$ , and that the stealthiness requirements remain consistent throughout the analysis, it follows that the evolution of the estimation error between two consecutive integrity enforcement blocks will depend only on  $E[\mathbf{e}_{t_1}^a] = \Delta \mathbf{e}_{t_1}$  and  $E[\mathbf{e}_{t_2}^a] = \Delta \mathbf{e}_{t_2}$ , or more specifically  $\mathcal{R}_{t_1}^\mu$  and  $\mathcal{R}_{t_2}^\mu$ . Thus, if  $\mathcal{R}_{t_2}^\mu \subseteq \mathcal{R}_{t_1}^\mu$  and  $\mathcal{R}_{1..t_2}^\mu \subseteq \mathcal{R}_{1..t_1}^\mu$ , then no new estimation error values can be reached after time  $t_2$  and the terminating time for the reachability analysis can be  $t_1$ , since after time  $t_2$  as well as after all following ends of integrity enforcement blocks the state estimation errors would start from a subset of the error values from  $\mathcal{R}_{t_1}^\mu$ . In addition, when the above terminating condition is satisfied, the global reachable region of the state estimation error can be obtained as  $\mathcal{R}^\mu = \bigcup_{k=1}^{\infty} \mathcal{R}_k^\mu = \bigcup_{k=1}^{t_1} \mathcal{R}_k^\mu = \mathcal{R}_{1..t_1}^\mu$ .

Consequently, using Algorithm 1 we can compute a periodic integrity enforcement policy that maximizes  $L$  (i.e., reduces the integrity enforcement rate) while limiting the attacker's influence. Specifically, the algorithm will result in the enforcement policy that ensures that the state of reachable estimation errors does not contain points outside the set of safe (i.e., acceptable) errors  $\mathcal{R}_{e^a}$ . In our evaluations in the case studies chapter, we define  $\mathcal{R}_{e^a}$  using a threshold  $\|\Delta \mathbf{e}_{max}\|_2$  for the maximal 2-norm of the expected state

---

**Algorithm 1** Procedure for design of periodic integrity enforcement policies.

---

**Inputs:** System model, safe reachable region  $\mathcal{R}_{e^a}$  for the state estimation error  $e^a$

```

1: Enforcement distance  $L = 0$ 
2: repeat
3:    $L = L + 1$ 
4:   Form policy  $(\mu, f, L)$  such that distance between consecutive elements in  $\mu$  is  $L$  and
      $t_0 = L$ 
5:   Assign  $t = 0$  and the reachable region  $\mathcal{R}_{1..t} = \emptyset$ 
6:   repeat
7:      $t_{old} = t$ 
8:      $\mathcal{R}_{1..t_{old}} = \mathcal{R}_{1..t}$ 
9:      $t = \min\{t' | t' \in \mu \wedge t' > t_{old}\}$ 
10:    Compute  $\mathbf{N}_{t_{old}+1}, \dots, \mathbf{N}_t, \mathbf{M}_{t_{old}+1}, \dots, \mathbf{M}_t$  from (5.5)
11:    Compute  $\Theta_t$  from (5.10)
12:    Compute  $\alpha(\varepsilon, tp, 2h + tp)$ 
13:    for  $k = t_{old} + 1, \dots, t$  do
14:      Compute  $\mathcal{R}_k^\mu$  using (5.21)
15:       $\mathcal{R}_{1..t} = \mathcal{R}_{1..t} \cup \mathcal{R}_k^\mu$ 
16:    end for
17:    until  $\mathcal{R}_{1..t} \subseteq \mathcal{R}_{1..t_{old}}$  and  $\mathcal{R}_t \subseteq \mathcal{R}_{t_{old}}$ 
18:    until  $\mathcal{R}_{1..t_{old}} \setminus \mathcal{R}_{e^a} \neq \emptyset$ 
19:  Accept policy  $(\mu, f, L - 1)$ 

```

---

estimation error due to attacks. Thus, the safety condition in Line 18 of the algorithm is mapped into  $\max(\|e_1^a\|_{max}, \dots, \|e_t^a\|_{max}) \geq \|\Delta e_{max}\|_2$ , where  $\|e_k^a\|_{max} = \max \|E[e_k^a]\|_2$  as computed in (5.22).

Finally, while we do not provide any guarantees that Algorithm 1 will always terminate, for all analyzed systems, including the case studies from the case study chapter, the condition in Line 17 was always eventually satisfied. Therefore, for all considered systems we have been able to use the algorithm to obtain periodic integrity enforcement policies that ensure desired estimation performance even in the presence of attacks.

## Cumulative Message Authentication

The most common cyber method to deal with MitM attacks is the use of MACs that ensure integrity of delivered sensor measurements. When standard MACs are employed, at every time-instance when integrity of a sensor measurement is enforced, the corresponding MAC for the measurement *only* is computed and attached to the transmitted packet. Thus, we can define standard intermittent integrity enforcement policies as in Definition 4.

Intuitively, a standard intermittent data authentication policy ensures that attack does not influence sensor values (i.e.,  $\mathbf{a}_k = \mathbf{0}$ ) within pre-specified time windows of  $f$  sample points; the end-points of these windows are captured in sequence  $\mu_s$  and all the windows are separated by at most  $L$  samples. In previous chapters, we showed that for a very general class of intrusion detectors, when a standard intermittent data authentication policy is used with  $f = \min(\psi, q_{un})$ , where  $\psi$  denotes the observability index of the  $(\mathbf{A}, \mathbf{C})$  pair and  $q_{un}$  denotes the number of unstable eigenvalues of  $\mathbf{A}$ , the attacker cannot introduce unbounded state estimation errors while remaining stealthy.

On the other hand, with *cumulative integrity enforcement policies*, blocks of  $f_c$  measurements are used to compute a single MAC (thus the name cumulative MACs), which is

then attached to communication packets containing last measurements from these blocks. Consequently, when a cumulative MAC computed over  $f_c$  last time points is received at time  $t_j$ , the controller will be able to detect whether false-data has been injected via MitM attacks in the last  $f_c$  transmissions. To formalize this notion, we start from the following definition.

**Definition 7.** *An intermittent cumulative data authentication policy  $(\mu_c, f_c, L_c)$  at time (i.e., sample point)  $k$ , such that  $\mu_c = \{t_j\}_{j=0}^{\infty}$  where  $t_0 \geq f_c$ , and for all  $j > 0$ ,  $t_{j-1} < t_j$ , with  $L_c = \sup_{j>0} (t_j - t_{j-1})$ , ensures that*

$$\mathbf{a}_t = \mathbf{a}_{t-1} = \dots = \mathbf{a}_{t-f_c+1} = \mathbf{0}, \forall t \in \mathcal{D}_k, \quad (6.1)$$

where  $\mathcal{D}_k = \{t_j \mid t_j \in \mu_c \text{ and } t_j \leq k\}$ .

Unlike when standard MACs are used as in Definition 4, cumulative data authentication policy denies attacker influence retroactively, only after a cumulative MAC computed over a block of  $f_c$  consecutive measurements, is received. Thus, if the attacker cannot reach a desired error by the time a cumulative MAC arrives, he should not insert false data during the time points used to compute the MAC; otherwise, modified data will not pass authentication when the MAC arrives, and the attack will be detected.

To illustrate this, consider Figure 6.1 that shows two attacks – one that attempts to stay stealthy only prior to time  $t_j = t + f_c$  ("Attack 1") and the other that remains stealthy after ("Attack 2"). Until time  $t_j$ , sensor data between  $t$  and  $t_j$  are unauthenticated, and thus as captured in (6.1) false-data vectors do not have to be zero. In this case, if  $\mathcal{R}_{risk}$  threshold is *Threshold 2* the attacker will drive system into  $\mathcal{R}_{risk}$ , and thus he does not need to be concerned with authentication at time  $t_j$ . On the other hand, if  $\mathcal{R}_{risk}$  threshold is *Threshold 1*, "Attack 1" is unable to reach  $\mathcal{R}_{risk}$  by  $t_j$ , and needs to continue after. However, with arrival of cumulative MAC at  $t_j$  the attack will be detected. Thus, in this case the attacker needs to adapt the strategy to "Attack 2", as this strategy remains stealthy even with newly considered integrity enforcement policy.

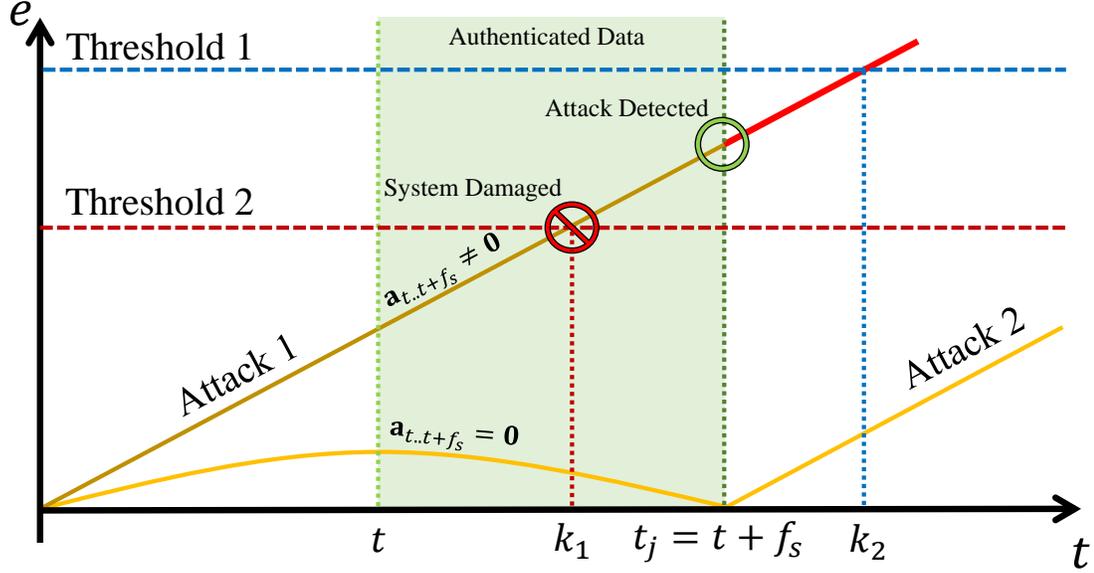


FIGURE 6.1: Stealthy attacks performed by the attacker, depending on the threshold of the  $\mathcal{R}_{risk}$  region. Cumulative MAC is assumed to arrive at  $t_j = t + f_c$ . When Threshold 1 describes  $\mathcal{R}_{risk}$  region, "Attack 1" is detected after cumulative MAC is received at  $t_j$  without reaching  $\mathcal{R}_{risk}$ , and thus attacker has to perform "Attack 2" to remain stealthy. On the other hand, when  $\mathcal{R}_{risk}$  boundary is Threshold 2, the attacker reaches  $\mathcal{R}_{risk}$  states before authentication, and successfully damages the system with "Attack 1" before it is detected.

To observe effects of cumulative authentication on the attacker, let us define attack vector's support set for a standard policy  $(\mu, f, L)$  according to Definition 4 as

$$\tilde{\mathcal{K}}_j = \begin{cases} \emptyset, & j - i \in \mu, \text{ for some } i, 0 \leq i < f, \\ \mathcal{K}, & \text{otherwise} \end{cases} \quad (6.2)$$

Then, from Definition 7, it follows that cumulative data authentication policy  $(\mu_c, f_c, L_c)$  at time  $k$  changes the support set of a stealthy attack  $\mathbf{a}_{1..k}$  as

$$\text{supp}(\mathbf{a}_j) = \begin{cases} \tilde{\mathcal{K}}_j, & \text{for some } j, j \leq k - f, \\ \mathcal{K}, & \text{otherwise} \end{cases} \quad (6.3)$$

Finally, for a stealthy attack  $\mathbf{a}_{1..k}$  from Definition 1, we denote its support set as  $\text{supp}(\mathbf{a}_{1..k}) = \mathcal{Q}_k \subseteq \{1, \dots, pk\}$ .

## 6.1 Reachability Analysis for Systems with Intermittent Cumulative Integrity Enforcements

Using the formal description of the attacker and cumulative integrity enforcement policies presented in previous section, we introduce a method to compute the set of all possible estimation errors caused by the attack that are obtainable with probability at least  $\eta$ . Parameter  $\eta$  exists to limit values of  $\mathbf{e}_k$  to realistically obtainable values, as  $\mathbf{e}_k$  is a Gaussian random variable and thus takes the values from an unbounded set. Set of obtainable estimation errors allows us to estimate effects of a stealthy attack on the system, as it shows deviation from steady state system states. We refer to this set as the  $k$ -reachable region of the state estimation error, and define it as in Definition 6

Following Lemma 4 we can obtain analytic solution for  $k$ -reachable regions under a specific cumulative data authentication policy, based on the following theorem.

**Theorem 16.** *The  $k$ -reachable region  $\mathcal{R}_k$  under an intermittent cumulative data authentication policy  $(\mu_c, f_c, L_c)$  can be overestimated as*

$$\overline{\mathcal{R}}_k = \left\{ \mathbf{e}_k^a \mid \mathbf{e}_k^a \mathbf{e}_k^{aT} \preceq \alpha_{\chi^2}^2 \mathbf{M}_k P_{\mathcal{Q}_k}^\dagger \Theta_k^{-1} (\mathbf{M}_k P_{\mathcal{Q}_k}^\dagger)^T + \gamma \Sigma \right\}, \quad (6.4)$$

where  $\alpha_{\chi^2}^2 = \alpha_{\chi^2}^2(\varepsilon, kp, 2h + kp)$  is an upper bound on  $\|\mathbf{z}_k^a\|_{\mathbf{Q}^{-1}}$ . Furthermore,

$$\Theta_k = \sum_{\tau=1}^k \mathbf{H}_\tau^T \mathbf{Q}^{-1} \mathbf{H}_\tau \quad (6.5)$$

$$\mathbf{H}_\tau = \left[ \mathbf{N}_\tau \mathbf{P}_{\mathcal{Q}_\tau}^\dagger \quad \mathbf{0}_{p \times (|\mathcal{Q}_k| - |\mathcal{Q}_\tau|)} \right] \quad (6.6)$$

$$\mathbf{N}_k = \left[ -\mathbf{C} \mathbf{A} \mathbf{M}_{k-1} \mid \mathbf{I} \right] \quad (6.7)$$

$$\mathbf{M}_k = - \left[ (\mathbf{A} - \mathbf{K} \mathbf{C} \mathbf{A})^{k-1} \mathbf{K} \mid \dots \mid \mathbf{K} \right]. \quad (6.8)$$

Similar results can be found in previous chapter and Kwon et al. (2015). However, there are two essential differences that separate these theorems. In previous chapter, data

is authenticated instantaneously, causing the attacker to plan ahead and providing limits even before authentication occurs. On the other hand, in Kwon et al. (2015), authors do not consider data authentication, and consider  $E[g_k^a] > h$  rather than  $P(g_k^a) > h$  as stealthiness condition.

*Proof.* From (5.7), it follows that

$$\beta_k = P(g_k > h) = P\left(\sum_{i=1}^k (\mathbf{z}_i^T \mathbf{Q}^{-1} \mathbf{z}_i) > 2h + kp\right).$$

Using the Lemma 4, the stealthiness condition can be overapproximated by

$$\|\Delta \mathbf{z}_k\|_{\mathbf{Q}^{-1}} \leq \alpha_{\chi^2}. \quad (6.9)$$

We represent equations from (2.10) and (2.10) in their non-recursive form, and substitute them in (6.9) to obtain that

$$\alpha_{\chi^2}^2(\varepsilon, kp, 2h + kp) - (\mathbf{P}_{\mathcal{Q}_k} \mathbf{a}_{1..k})^T \Theta_k \mathbf{P}_{\mathcal{Q}_k} \mathbf{a}_{1..k} \geq 0 \quad (6.10)$$

needs to be satisfied in order for the attacker to remain stealthy. From Lemma 13,  $\Theta_k$  is positive definite, and we can form Schur complement as

$$\begin{bmatrix} \Theta_k^{-1} & \mathbf{P}_{\mathcal{Q}_k} \mathbf{a}_{1..k} \\ (\mathbf{P}_{\mathcal{Q}_k} \mathbf{a}_{1..k})^T & \alpha_{\chi^2}^2(\varepsilon, kp, 2h + kp) \end{bmatrix} \geq 0. \quad (6.11)$$

In order to generalize error computation for any stealthy attack  $\mathbf{a}_{1..k}$ , let us introduce matrix

$$\mathbf{G} = \begin{bmatrix} -\mathbf{M}_k \mathbf{P}_{\mathcal{Q}_k}^\dagger & \mathbf{0}_{n \times 1} \\ \mathbf{0}_{1 \times |\mathcal{Q}_k|} & 1 \end{bmatrix} \quad (6.12)$$

and form quadratic representation

$$\mathbf{G} \begin{bmatrix} \Theta_k^{-1} & \mathbf{P}_{\mathcal{Q}_k} \mathbf{a}_{1..k} \\ (\mathbf{P}_{\mathcal{Q}_k} \mathbf{a}_{1..k})^T & \alpha_{\chi^2}^2(\varepsilon, kp, 2h + kp) \end{bmatrix} \mathbf{G}^T \geq 0, \quad (6.13)$$

from which follows

$$\begin{bmatrix} \mathbf{M}_k \mathbf{P}_{\mathcal{Q}_k}^\dagger \boldsymbol{\Theta}_k^{-1} (\mathbf{M}_k \mathbf{P}_{\mathcal{Q}_k}^\dagger)^T & -\mathbf{M}_k \mathbf{P}_{\mathcal{Q}_k}^\dagger \mathbf{P}_{\mathcal{Q}_k} \mathbf{a}_{1..k} \\ -(\mathbf{P}_{\mathcal{Q}_k} \mathbf{a}_{1..k})^T (\mathbf{M}_k \mathbf{P}_{\mathcal{Q}_k}^\dagger)^T & \alpha_{\chi^2}^2(\varepsilon, kp, 2h + kp) \end{bmatrix} \succeq 0. \quad (6.14)$$

Once again, we use Schur complement and non-recursive form of (2.10) and (2.10), to obtain

$$\mathbf{M}_k P_{\mathcal{Q}_k}^\dagger \boldsymbol{\Theta}_k^{-1} (\mathbf{M}_k P_{\mathcal{Q}_k}^\dagger)^T - \frac{1}{\alpha_{\chi^2}^2} \Delta \mathbf{e}_k \Delta \mathbf{e}_k^T \succeq 0. \quad (6.15)$$

Given that  $\Delta \mathbf{e}_k$  is deterministic signal from (2.10) and (2.10),

$$\Delta \mathbf{e}_k = E(\Delta \mathbf{e}_k) = E(\mathbf{e}_k^a) - E(\mathbf{e}_k) = E(\mathbf{e}_k^a) \quad (6.16)$$

it follows that first condition in  $\mathcal{R}_k$  is overapproximated by

$$E[\mathbf{e}_k^a] E[\mathbf{e}_k^a]^T + \gamma \boldsymbol{\Sigma} \leq \alpha_{\chi^2}^2 \mathbf{M}_k P_{\mathcal{Q}_k}^\dagger \boldsymbol{\Theta}_k^{-1} (\mathbf{M}_k P_{\mathcal{Q}_k}^\dagger)^T + \gamma \boldsymbol{\Sigma}.$$

Our initial assumption was (6.9), which generates attack set  $\overline{\mathcal{A}}_k$  from Lemma 4 that overestimates  $\mathcal{A}_k$ . Thus, the second condition in  $\mathcal{R}_k$  that  $\mathbf{a}_{1..k} \in \mathcal{A}_k \subseteq \overline{\mathcal{A}}_k$  is satisfied, which concludes the proof.  $\square$

If we denote  $\mathbf{Y} = \alpha_{\chi^2}^2 \mathbf{M}_k P_{\mathcal{Q}_k}^\dagger \boldsymbol{\Theta}_k^{-1} (\mathbf{M}_k P_{\mathcal{Q}_k}^\dagger)^T + \gamma \boldsymbol{\Sigma}$ , then by using Schur complement one more time, overapproximation of k-reachable region  $\mathcal{R}_k$  becomes

$$\overline{\mathcal{R}}_k = \left\{ \mathbf{e}_k^a \mid \mathbf{e}_k^{aT} \mathbf{Y}^{-1} \mathbf{e}_k^a \leq 1 \right\}, \quad (6.17)$$

defined by inequality that describes an ellipsoid, and can be easily computed or even determined analytically.

Finally, we deem cumulative data authentication policy  $(\mu_c, f_c, L_c)$  to be successful when reachable estimation error region does not achieve some high-risk region  $\mathcal{R}_{risk}$  that causes the system to malfunction. Formally,

$$(\forall k \in \mathbb{N}) \quad \overline{\mathcal{R}}_k \cap \mathcal{R}_{risk} = \emptyset.$$

A policy that satisfy this condition always exists (e.g., when  $f_c = L_c = 1$ ). Furthermore, we obtain the following result using the approach from previous chapter to show a similar result for systems where standard MACs are intermittently employed; the idea is to map the problem into the problem of bounding the reachable set when standard intermittent integrity enforcement policy  $(\mu_c, f_c, L_c + f_c)$  is used.

**Theorem 17.** *Consider an LTI system from (2.10) and (2.10) with an intermittent cumulative data integrity policy  $(\mu_c, f_c, L_c)$ , where*

$$f_c = \min(\psi, q_{un}), \quad (6.18)$$

*$L_c$  is finite,  $\psi$  is the observability index of the  $(\mathbf{A}, \mathbf{C})$  pair, and  $q_{un}$  denotes the number of unstable eigenvalues of  $\mathbf{A}$ . Then  $(\forall k \in \mathbb{N}) \mathcal{R}_k$  is bounded.*

Proof of the theorem follows from previous chapters. Although from theoretical standpoint an infinite number of steps needs to be explored, as  $\mathcal{R}_k$  will converge to its bounds arbitrarily slow, in practical tests we observed that after some system-dependent time point  $k_t$ ,  $\overline{\mathcal{R}_{k_t+1}} \subseteq \cup_{k=1}^{k_t} \overline{\mathcal{R}_k}$  which terminates the search. Thus, we obtain a full procedure to design safe cumulative integrity enforcement policies as presented in Algorithm 2.

---

**Algorithm 2** Procedure for deriving periodic integrity enforcement policies with cumulative  $f$  steps authentication.

---

**Inputs:** System model, high-risk set of states  $\mathcal{R}_{risk}$ , number of consecutively authenticated messages  $f$ , additional probability of the detection introduced by the attacker  $\varepsilon$ .

- 1: Enforcement distance  $L = 0$
  - 2: **repeat**
  - 3:    $L = L + 1$
  - 4:   Form policy  $(\mu, f, L)$  as in Theorem 17 with  $t_0 = L$
  - 5:   Assign  $k = 0$ , union of reachable regions  $\mathcal{R}_\cup = \emptyset$ , and  $\overline{\mathcal{R}}_0 = \mathbf{0}$
  - 6:   **repeat**
  - 7:      $\mathcal{R}_\cup = \mathcal{R}_\cup \cup \overline{\mathcal{R}}_k$
  - 8:      $k = k + 1$
  - 9:     Compute  $\mathbf{N}_k$  and  $\mathbf{M}_k$  from (6.7) and (6.8)
  - 10:    Compute  $\Theta_k$  from (6.5)
  - 11:    Compute  $\alpha(\varepsilon, kp, 2h + kp)$
  - 12:    Compute  $\overline{\mathcal{R}}_k$  from (6.17)
  - 13:    **until**  $\overline{\mathcal{R}}_k \subseteq \mathcal{R}_\cup$
  - 14: **until**  $\overline{\mathcal{R}}_k \cap \mathcal{R}_{risk} \neq \emptyset$
  - 15: Accept policy  $(\mu, f, L - 1)$
-

## Case Studies

In this chapter we use automotive case studies to illustrate how intermittent data integrity enforcements can ensure satisfiable control performance even in the presence of attacks. For all studies, sensor values are transmitted over an internal vehicle’s network, such as commonly used CAN bus. Note that in [Lesi et al. (2017a)], we provide additional automotive case-studies (and the overall scheduling framework) for intermittent authentication of CAN-bus messages from system sensors, and in [Lesi et al. (2017b)] we show benefits of intermittent authentication on vehicle’s ECU scheduling.

### 7.1 Case Study: Vehicle Trajectory Following

We start with the model used in [Kerns et al. (2014)] to describe vulnerabilities and potential attacks on autonomous systems adapted for two-axis tracking; we obtain the following discretized models (with sampling period of  $0.01s$ ) for each axis

$$\mathbf{A}_d = \begin{bmatrix} 1 & 0.01 \\ 0 & 1 \end{bmatrix} \quad \mathbf{B}_d = \begin{bmatrix} 0.0001 \\ 0.01 \end{bmatrix} \quad \mathbf{C}_d = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (7.1)$$

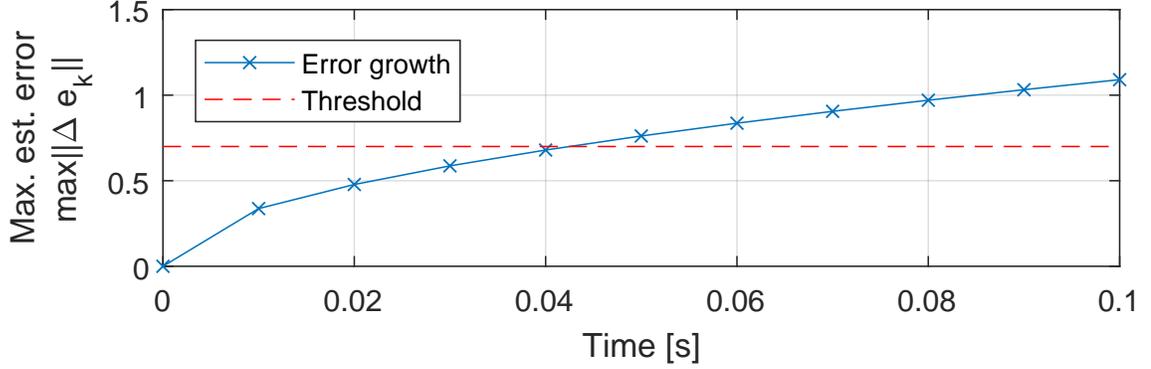


FIGURE 7.1: Evolution of the maximal estimation error for vehicle tracking; without integrity enforcements, the attacker forces the system outside of the safe range in 4 steps.

Assume that the attacker can modify the values from all sensors. The system is perfectly attackable as the matrix  $\mathbf{A}_d$  is unstable and  $\text{supp}(\mathbf{C}\mathbf{v}) \in \mathcal{K}$ , since  $\mathcal{K} = \mathcal{S}$ .

We consider the largest additive estimation error on position to be  $0.5 \text{ m}$  and on speed to be  $0.5 \frac{\text{m}}{\text{s}}$ , resulting in  $\|\Delta \mathbf{e}_{max}\|_2 = 0.7$ . We also set such that the probability of false positive from (2.8) to  $\beta = 1.5\%$ , and additional probability of detection introduced by the attacker from (2.16) to  $\varepsilon = 0.1\%$ .

Without integrity enforcements, the attacker could force the state estimation error above  $\|\Delta \mathbf{e}_{max}\|_2$  threshold after 4 steps, as shown in Figure 7.1. We considered three periodic integrity enforcement policies with  $f = 1$  as specified in conditions of Theorem 7, and periods  $L = 20, 30$  and  $35$ , denoted by  $\mu_{20}, \mu_{30}$ , and  $\mu_{35}$  respectively. Using results from Chapter 5, we show that the first two policies are safe, while the third policy can violate the  $\|\Delta \mathbf{e}_{max}\|_2$  threshold – Figure 7.2 illustrates the evolution of the maximal estimation errors for each policy.

Finally, we evaluate the effects of intermittent integrity guarantees for trajectory following on a circular path with  $100 \text{ m}$  radius, at speed of  $3.14 \frac{\text{m}}{\text{s}}$ . Figure 7.3 shows results of  $200 \text{ s}$  long simulations, with attacks starting at  $100 \text{ s}$ . As illustrated, when integrity is enforced on less than  $3.4\%$  of messages, i.e. when  $\mu_{30}$  is employed, we have strong control performance guarantees in the presence of attacks on all vehicle sensors.

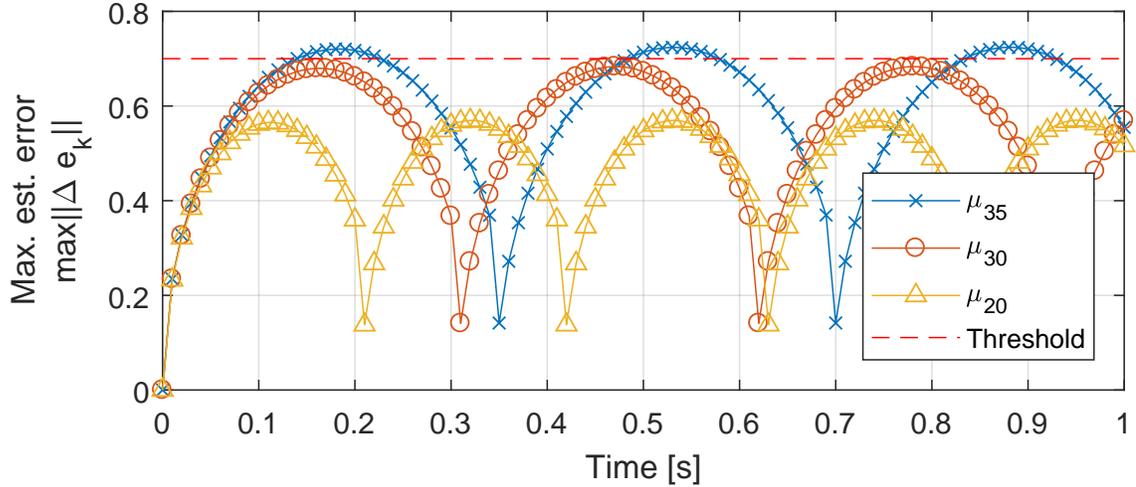
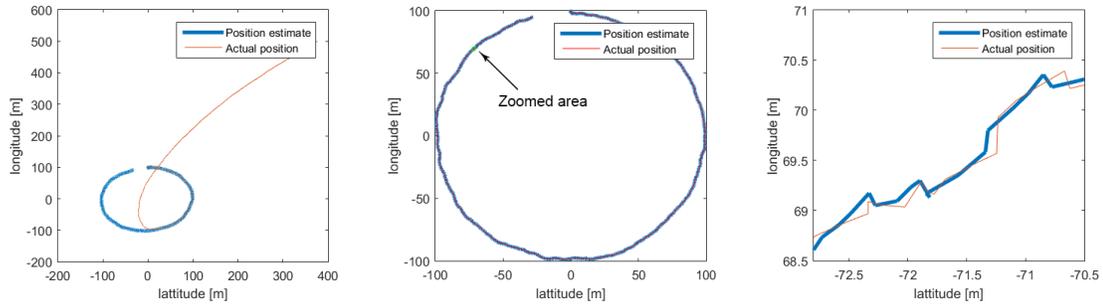


FIGURE 7.2: Maximal estimation error in the presence of attacks on all sensors for vehicle-tracking case study with three different integrity enforcement policies with  $f = 1$  and periods  $L = 20, 30, 35$ .



(a) State Estimates for system under stealthy attack, without integrity enforcement policies (b) State estimates under stealthy attack with integrity enforcement policies (c) Zoomed section of the Fig. 7.3(b).

FIGURE 7.3: State estimation of the tracked vehicle trajectory - without integrity enforcements a stealthy attacker can introduce a significant estimation error in a short period of time. However, even with intermittent integrity enforcement, the attack effects are negligible. Duration of the simulation is 200 s and the attack starts at 100 s.

## 7.2 Degraded Cooperative Adaptive Cruise Control (dCACC)

Cooperative Adaptive Cruise Control (CACC) employs communication to obtain smaller following distance and better platooning stability than standard Adaptive Cruise Control. To achieve this, each vehicle is equipped with lidar and acceleration measurement sent

from the preceding vehicle. However, when acceleration data is not available CACC needs to switch to dCACC, that is based only on local vehicle measurements. In this mode, Singer acceleration model is used to estimate acceleration of the preceding vehicle [Ploeg et al. (2015a)] – i.e.,

$$\begin{bmatrix} \dot{d} \\ \dot{v} \\ \dot{a} \end{bmatrix} = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -\frac{1}{\tau} \end{bmatrix} \begin{bmatrix} d \\ v \\ a \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} [u] \quad (7.2)$$

$$\mathbf{y} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} d \\ v \\ a \end{bmatrix}. \quad (7.3)$$

Here,  $d$  denotes the distance of the vehicle from the preceding vehicle,  $v$  is its speed – both computed from lidar measurements and transmitted over the bus,  $a$  is the acceleration,  $u$  is the control input, while  $\tau = 0.8$  represents maneuver time constant of the preceding vehicle [Ploeg et al. (2015a)]. We focus on the cases when the attacker compromises all car sensors, making the system perfectly attackable. We set maximal estimation error to be  $0.5m$  on position,  $3.3\frac{m}{s}$  on speed, and  $0.3\frac{m}{s^2}$  on acceleration, resulting in  $\|\Delta\mathbf{e}_{max}\|_2 = 3.351$ .

As in trajectory tracking, we assume  $\varepsilon = 0.1\%$ , and  $\beta = 0.35\%$ . Since observability index  $\psi = 2$  and number of unstable eigenvalues of  $\mathbf{A}$  is 2, then  $f = 2$ . For periodic policy with  $L = 20$  we obtain the maximal reachable estimation errors in the presence of stealthy attacks as presented in Figure 7.4. In addition, visual representation of reachable regions with this policy in comparison to a system without integrity enforcement is shown in Figure 7.5. These results illustrate that even with 10% authenticated messages the system ensures satisfiable performance under false-date injection attacks.

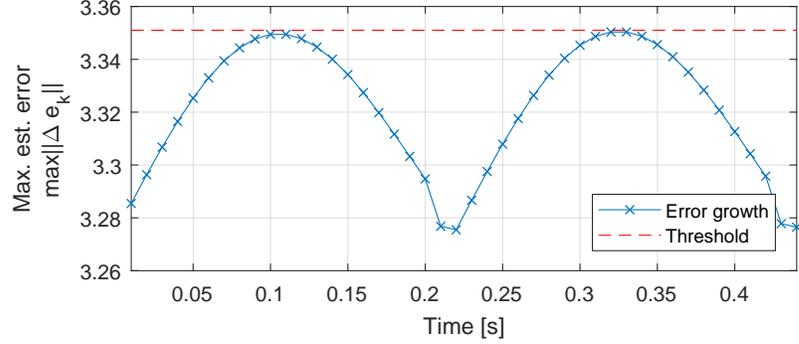


FIGURE 7.4: Evolution of maximal estimation error for dCACC. If we can enforce integrity on two sensor values after every twenty unsecured sensor values, the system remains under the specified safety threshold.

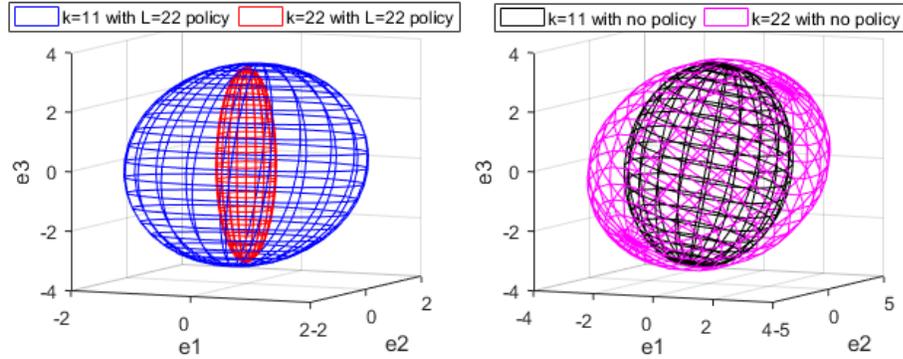


FIGURE 7.5: Reachable state estimation errors in the presence of stealthy attacks for dCACC in steps  $k = 11$  and  $k = 22$  with and without data integrity enforcement. Without integrity enforcement, the size of reachable regions keeps increasing, while when integrity is being enforced with policy  $L = 20$  and  $f = 2$ , estimation error evolves as in Figure 7.4, and the attacker is contained between red and blue ellipsoids.

### 7.3 Case Study: Delayed Authentication

We demonstrate our method on a steering control case study [Rajamani (2011)]. We consider a vehicle that weighs  $m = 1573kg$ , and measures  $l_f = 1.1m$  in front and  $l_r = 1.58m$  behind its center of gravity. Yaw moment of inertia is  $I_z = 2873kgm^2$ , and front and rear cornering coefficients are both  $C_{af} = C_{ar} = 80k$ . At the time of the attack, vehicle is assumed to be running at constant speed of  $v_x = 35m/s$ . Furthermore, we assume the system to be equipped with Kalman filter, a controller, and intrusion detector. Probability

of the false positive is set to 5%, and additional probability introduced by the attacker is assumed as  $\varepsilon = 0.01\%$ . The model of the vehicle is provided as in (2.1), where:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -2\frac{C_{af}C_{ar}}{mv_x} & 2\frac{C_{af}C_{ar}}{m} & 2\frac{-C_{af}l_f+C_{ar}l_r}{mv_x} \\ 0 & 0 & 0 & 1 \\ 0 & -2\frac{C_{af}l_f-C_{ar}l_r}{I_zv_x} & 2\frac{C_{af}l_f-C_{ar}l_r}{I_z} & -2\frac{C_{af}l_f^2+C_{ar}l_r^2}{I_zv_x} \end{bmatrix} \quad (7.4)$$

$$\mathbf{B} = \begin{bmatrix} 0 & 2C_{af}/m & 0 & 2\frac{C_{af}l_f}{I_z} \end{bmatrix}^T; \quad \mathbf{C} = \mathbf{I}_4 \quad (7.5)$$

We discretized system matrices to obtain the model implemented in the car. As  $\mathcal{R}_{risk}$  set, we choose an inverted hyperball in  $\mathbb{R}^4$  space, centered at  $\mathbf{0}$ , and of diameter  $e_{max}$ . Thus, the attack is successful when  $\|\mathbf{e}_k\|_2 \geq e_{max}$ . Specifically, we allow the error in states to be 0.2 m on lateral position, 0.5 m/s on lateral speed, 0.3 rad for steering angle, and 0.3 rad/s for angular velocity of axle, which results in  $e_{max} = 0.6856$ . Using the conditions from [Mo et al. (2010); Jovanov and Pajic (2017b)], it can be shown that  $\lim_{k \rightarrow \infty} \mathcal{R}_k$  is unbounded set, when no data integrity enforcement policy is used, and thus the attacker will be able to reach  $\mathcal{R}_{risk}$ .

To protect from the attack, we choose a periodic cumulative integrity enforcement policy – i.e., enforcement occurs at equidistant intervals  $L_c$ . Furthermore, we considered the case where authentication message is sent at the end of the authentication block taking equal bandwidth for any number of authenticated samples. Thus, we allow  $f_c = L_c$ , effectively reducing the attack to  $\mathbf{0}$  at the end of the authentication block. We tried different values of  $L_c$  until we reached  $L_c = 5$  for which the attacker could reach  $\mathcal{R}_{risk}$ . Thus, the highest possible reduction in resources used for authentication while satisfying safety constraints will be  $L_c = 4$ . Plot of the maximal error norm evolution for case without authentication versus the case with periodic authentication parameterized by  $L_c = f_c = 4$  is shown in Figure 7.6. In addition, Figure 7.7 shows a 4s simulation of possible evolutions of estimation error over time if the system is compromised at 2s, and  $L_c = f_c = 4$  is

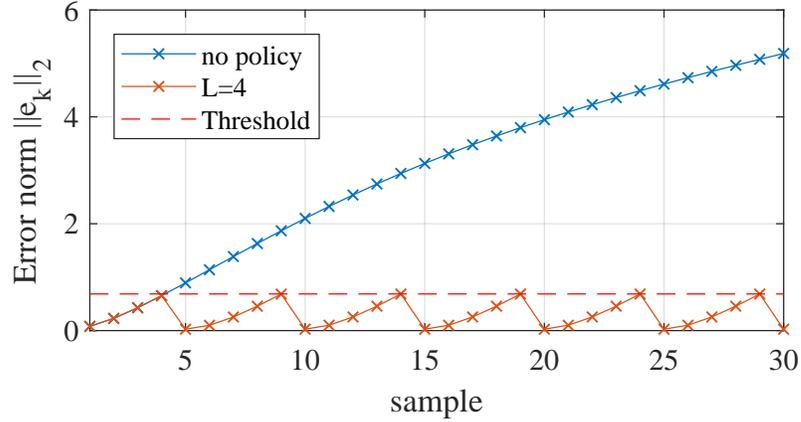


FIGURE 7.6: Evolution of the norm of estimation error. Figure shows two system realizations – one that uses unauthenticated data, and the other that authenticates previous  $f = 4$  points of data every  $L = 4$  samples. Authentication policy defined like this retains the estimation error inside of the safe states.

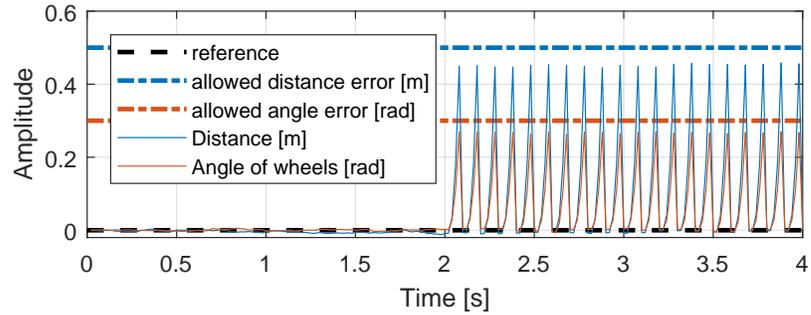


FIGURE 7.7: A 4s simulation of the system, with the attack gaining access to the sensors at 2s. Figure shows an error that the attacker could induce prior to being detected when authentication arrives. This figure shows possible attacks that depend on when the attacker injects erroneous values, and thus only one of the peaks shown in the figure could be reached prior to detection of the attack.

employed.

# 8

## Implementation

For practical implementation of our theory we used a system originally developed at University of Waterloo, comprised of three battery-powered electrical-motor hobby cars moving on a treadmill, shown in Figure 8.1. Videos of system under the attacks can be seen on <https://youtu.be/E3NT1DWqsCA> and <https://youtu.be/1-Nw9bKfF9Y>. Each of the cars is equipped with wireless transmitters/receivers, microcontrollers, origi-



FIGURE 8.1: Snapshot of the car platoon system used for tests.

nal speed and steering actuators, battery current sensors, and AprilTags – black and white check-board patterns on a paper attached to the top of the car. Wireless receivers and microcontrollers process the inputs sent from a computer and control the actuators, trans-

mitters send information about recorded battery current, while AprilTags get identified by a camera positioned above the treadmill, allowing for heading and position measurements. These measurements are sent to the computer executing robot operating system (ROS), and filtered through One Euro Filter. ROS uses these measurements to simulate global positioning system (GPS) and lidar sensors on the cars. These measurements are processed in ROS and used to close the loop that regulates car platoon behavior. In addition to camera measurements, ROS also acquires measurements from rotary encoder mounted on treadmill for further signal processing and control. All sensors acquire information and system operates at  $30Hz$ . Overview of this system is shown in Figure 8.2.

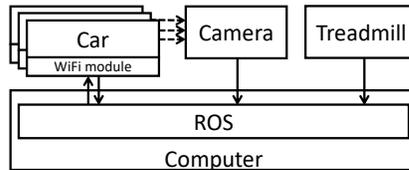


FIGURE 8.2: Full platooning system, with cars moving on top of a treadmill. ROS sends steering and speed commands to cars over the wireless network. Position and heading of cars is then recorded by a camera with help of AprilTags, and sent back to ROS through camera interface. ROS utilizes this data to compute corrected steering and speed that it will send to the cars.

Control of the car on ROS is implemented as a composition of two controllers, lateral and longitudinal. Lateral controller is Stanley controller that utilizes geometric bicycle model [Snider et al. (2009)]. Longitudinal controller is comprised of two cascaded PID controllers, as shown in Figure 8.3 <sup>1</sup>. Simulated lidar detected nearest cars using 360 raycast from car's position.

Although architecture of the system is thoroughly described, Waterloo platooning system lacked a model and intrusion detection mechanisms. Given that Kalman estimator is required for  $\chi^2$ , CUSUM, and SPRT detectors described in previous chapters, and system model is required for Kalman filter implementation, we provide system identification first.

<sup>1</sup> Figure obtained from David Donghyn Shin's (University of Waterloo) master thesis.

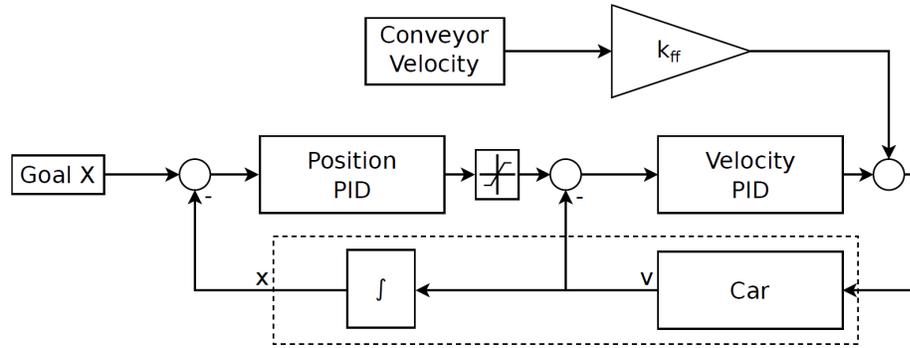


FIGURE 8.3: Longitudinal movement controller for a vehicle in the vehicle platoon. Desired velocity in relation to camera is 0 once the goal position is achieved, and thus we include conveyor velocity.

## 8.1 Car and Platoon System Identification

Prior to identifying the system, we fully charged car batteries and let the cars run for  $30min$  to avoid non-linearities related to battery drain. Prior experiments showed that this nonlinearity does not significantly affect system behavior and can be included in system noise component. We considered position commanded to the lead vehicle to be the input to the system as all other vehicles are controlled in relation to the lead. Outputs recorded for the sake of system identification were recorded position of the cars. Input signals were designed such that we have a good idea of nonlinearities related to the changes in the amplitude of the input, versus the change of the amplitude of the output as shown in Figure 8.4. First, we performed a short train of step inputs with increasing amplitude to assess the ratio between inputs and outputs for different values. Next, we recorded a train of longer step inputs to record potential changes in rise time and steady state. We also recorded an increasing and decreasing stair-like signal to identify nonlinearities due to changes during different operating points. Finally, we introduced random inputs to evaluate identified system. Only long step inputs were used in system identification algorithms, and the rest of the signal was used for non-linearity evaluation and testing. Identification was performed using Subspace System Identification method, with original

codes being developed by professor Henry Gavin.

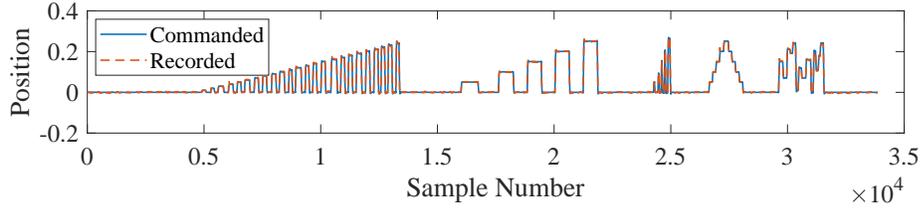


FIGURE 8.4: Signal used for system identification along with the outputs of the system. Horizontal axis is provided in samples as subspace system identification returns system in discrete time.

Position of the car is recorded and commanded in relation to the camera. This means that when treadmill and car speeds match, position of the car does not change, leading to 0 change in position for non-zero speed of the car. Thus, when identifying the system, either 1) speed of the treadmill needs to be accounted for as distance covered by the car, or 2) speed of the treadmill needs to be deducted from the speed of the car. In our current system identification we used second option. Furthermore, we simplified identification by observing only longitudinal movement as it was sufficient to evaluate performance of the system under attacks. Longitudinal dynamics were of greater interest compared to lateral, as any inconsistencies in control would cause cars to crash.

Initially, we tried identifying longitudinal dynamics of a single car. Most system identification algorithms assume that inputs are independent from the outputs, which is not the case in the current vehicle platoon setup. For a system set up in a way where we directly control the throttle of the car, issue of car flying off the treadmill came up. This made it impossible to detect dynamics of a car moving on a surface without positional controller, so we decided to identify entirety of the system in Figure 8.3 as a single system block, using commanded position as an input and measured position as an output. In addition to this, we also identified the whole platoon as a single system, having commanded position of the first car as the input and recorded positions of all three cars as the output. Obtained

system matrices for a single car were as follows.

$$\mathbf{A} = \begin{bmatrix} 0.9739 & 0.1729 & 0.2425 & 0.5519 & 1.2201 & 0.8052 & 1.1773 & 1.2174 \\ -0.0302 & 0.8931 & -0.2171 & -0.3140 & -0.6540 & -0.3800 & -0.6404 & -0.5565 \\ 0.0014 & 0.0148 & 0.9603 & -0.2736 & -0.3778 & -0.4833 & -0.2928 & -0.7008 \\ 0.0023 & 0.0172 & 0.0156 & 0.6836 & -1.0222 & -0.3414 & -0.7206 & -0.2290 \\ -0.0000 & 0.0002 & -0.0033 & 0.0457 & 0.8467 & -0.4383 & -0.1960 & -0.4851 \\ -0.0000 & -0.0000 & 0.0005 & 0.0025 & 0.0334 & 0.9867 & -0.3511 & 0.0433 \\ -0.0000 & -0.0003 & 0.0030 & 0.0111 & -0.0020 & 0.3362 & 0.7640 & -0.5923 \\ -0.0000 & -0.0001 & 0.0011 & 0.0029 & 0.0148 & -0.0168 & 0.0712 & 0.9221 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 0.0001 & 0.0463 & -0.0130 & -0.0144 & 0.0014 & 0.0002 & 0.0006 & 0.0001 \end{bmatrix}^T$$

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{D} = \begin{bmatrix} -2.5532e - 05 \end{bmatrix}$$

While a single car can usually be described by two states (position and velocity), there existed a delay of three samples between inputs and outputs, causing increase in states to account for delayed samples. As for the vehicle platoon case, there was a total of eight states as well, as identified by the subspace identification method. This corresponds to the theoretical case [Ploeg et al. (2015b)] where leading car has two states (velocity and acceleration), while following cars have three states (distance from preceding vehicle, velocity, and acceleration). Matrices of the platoon are as follows.

$$\mathbf{A} = \begin{bmatrix} 0.9800 & -0.0350 & -0.0152 & -0.0304 & -0.0106 & 0.0154 & -0.0014 & 0.0008 \\ 0.0310 & 0.9777 & -0.0600 & -0.0280 & -0.0126 & -0.0427 & -0.0038 & 0.0170 \\ -0.0140 & 0.0565 & 0.9643 & -0.0585 & 0.0728 & 0.0016 & 0.0339 & 0.0216 \\ 0.0092 & -0.0043 & 0.0551 & 0.9598 & -0.0291 & -0.0017 & 0.0528 & 0.0656 \\ -0.0102 & 0.0217 & -0.0586 & 0.0570 & 0.9166 & 0.0077 & -0.0925 & 0.0828 \\ -0.0056 & 0.0150 & -0.0124 & 0.0038 & -0.0454 & 0.9366 & 0.0794 & 0.0064 \\ 0.0044 & -0.0073 & 0.0124 & -0.0262 & 0.0651 & 0.0115 & 0.9236 & -0.0851 \\ -0.0021 & 0.0043 & -0.0043 & -0.0097 & -0.0223 & -0.0516 & 0.1207 & 0.9154 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} -0.0535 & 0.2523 & -0.4824 & 0.4638 & -0.9999 & -0.7903 & 0.6750 & -0.5782 \end{bmatrix}^T$$

$$\mathbf{C} = \begin{bmatrix} -0.0440 & 0.0375 & -0.0157 & -0.0069 & 0.0163 & 0.0251 & 0.0021 & -0.0034 \\ -0.0692 & 0.0003 & 0.0245 & -0.0091 & -0.0272 & 0.0122 & -0.0112 & 0.0053 \\ -0.0797 & -0.0518 & -0.0363 & -0.0325 & 0.0010 & -0.0057 & 0.0072 & 0.0101 \end{bmatrix}$$

$$\mathbf{D} = \begin{bmatrix} 0.0186 & 0.0066 & 0.0024 \end{bmatrix}^T$$

## 8.2 Estimation and Intrusion Detection

Given these system matrices, we implemented Kalman state estimator for each of the cars and entire platoon. While this estimator was not used to control the system, we utilized obtained residual to detect any anomalies or attacks. To test the system performance, we designed several attacks: Simple step attack where attacker does not take IDS into account, slow ramp attack that tries to avoid detection, more aggressive ramp attack that is more likely to be detected, and stealthy attacks as described in previous chapters that have 0 attack at the time of intermittent integrity enforcement, where we varied aggressiveness of the attacker and duration between consecutive integrity enforcements. All the attacks affected only middle car of the platoon, as all mistakes in position for this car resulted in crash. Traces of residual and system states can be observed in Figure 8.5 and Figure 8.6 respectively for a single attacked car, and in Figure 8.7 and Figure 8.8 respectively for the whole platoon. As can be observed, ramp attacks in our system are very difficult to identify and can cause significant damage. However, when intermittent integrity enforcement is implemented, if the attacker introduces significant change in states, residual spikes and the attacker is detected.

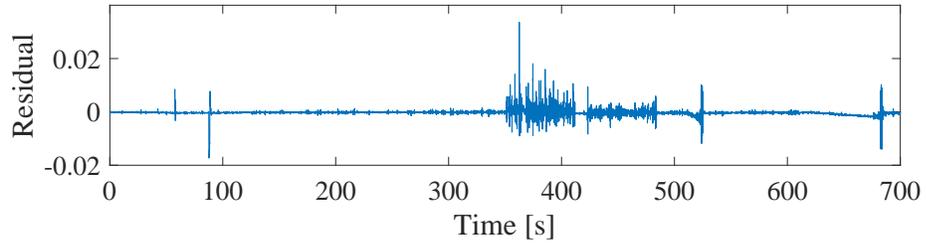


FIGURE 8.5: Residuals obtained from one-car Kalman filter during the attacks. We can see residual spiking for a high-risk high-influence attack.

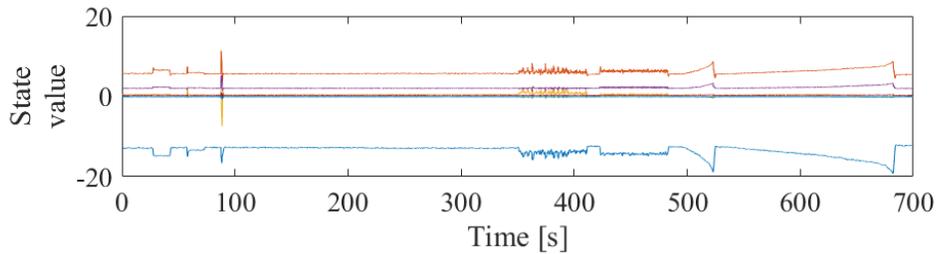


FIGURE 8.6: States of a single car. We can see that ramp attacks cause significant change in states without significant effect on residual.

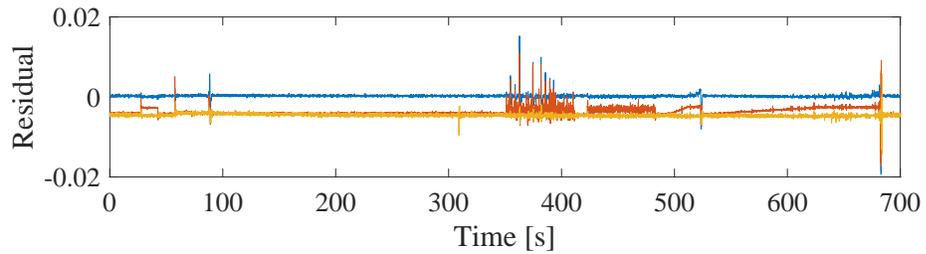


FIGURE 8.7: Residuals obtained from whole platoon Kalman filter during the attacks. We can see residual spiking for a high-risk high-influence attack.

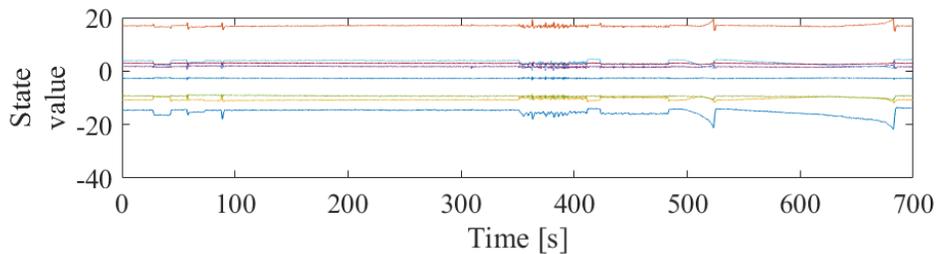


FIGURE 8.8: States of a three-car platoon. We can see that ramp attacks cause significant change in states without significant effect on residual.

# Bibliography

- Amin, S., Schwartz, G. A., Cardenas, A. A., and Sastry, S. S. (2015), “Game-Theoretic Models of Electricity Theft Detection in Smart Utility Networks: Providing New Capabilities with Advanced Metering Infrastructure,” *IEEE Control Systems*, 35, 66–81.
- Boyd, S. and Vandenberghe, L. (2004), *Convex optimization*, Cambridge university press.
- Checkoway, S., McCoy, D., Kantor, B., Anderson, D., Shacham, H., Savage, S., Koscher, K., Czeskis, A., Roesner, F., Kohno, T., et al. (2011), “Comprehensive experimental analyses of automotive attack surfaces,” in *Proceedings of USENIX Security*.
- Fawzi, H., Tabuada, P., and Diggavi, S. (2014), “Secure Estimation and Control for Cyber-Physical Systems Under Adversarial Attacks,” *IEEE Transactions on Automatic Control*, 59, 1454–1467.
- Granjon, P. (2014), “The CUSUM algorithm a small review,” .
- Johnson, N., Kotz, S., and Balakrishnan, N. (1995), *Continuous univariate distributions*, Wiley & Sons.
- Jovanov, I. and Pajic, M. (2017a), “Relaxing Integrity Requirements for Resilient Control Systems,” *CoRR*, abs/1707.02950.
- Jovanov, I. and Pajic, M. (2017b), “Sporadic Data Integrity for Secure State Estimation,” in *56th IEEE Conference on Decision and Control (CDC)*, pp. 163–169.
- Juang, J.-N. (1994), *Applied System Identification*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA.
- Kerns, A. J., Shepard, D. P., Bhatti, J. A., and Humphreys, T. E. (2014), “Unmanned aircraft capture and control via GPS spoofing,” *Journal of Field Robotics*, 31, 617–636.
- Kwon, C. and Hwang, I. (2017), “Reachability Analysis for Safety Assurance of Cyber-Physical Systems against Cyber Attacks,” *IEEE Transactions on Automatic Control*, PP, 1–1.
- Kwon, C., Liu, W., and Hwang, I. (2014), “Analysis and Design of Stealthy Cyber Attacks on Unmanned Aerial Systems,” *Journal of Aerospace Information Systems*, 1.

- Kwon, C., Yantek, S., and Hwang, I. (2015), “Real-Time Safety Assessment of Unmanned Aircraft Systems Against Stealthy Cyber Attacks,” *Journal of Aerospace Information Systems*, pp. 1–19.
- Langner, R. (2011), “Stuxnet: Dissecting a cyberwarfare weapon,” *Security & Privacy, IEEE*, 9, 49–51.
- Lesi, V., Jovanov, I., and Pajic, M. (2017a), “Network Scheduling for Secure Cyber-Physical Systems,” in *IEEE Real-Time Systems Symposium (RTSS)*.
- Lesi, V., Jovanov, I., and Pajic, M. (2017b), “Security-Aware Scheduling of Embedded Control Tasks,” *ACM Trans. Embed. Comput. Syst.*, 16, 188:1–188:21.
- Lin, C.-W., Zheng, B., Zhu, Q., and Sangiovanni-Vincentelli, A. (2015), “Security-Aware Design Methodology and Optimization for Automotive Systems,” *ACM Trans. on Des. Autom. of Elec. Syst.*, 21, 18.
- Lun, Y. Z., D’Innocenzo, A., Malavolta, I., and Benedetto, M. D. D. (2016), “Cyber-Physical Systems Security: a Systematic Mapping Study,” *CoRR*, abs/1605.09641.
- Miao, F., Pajic, M., and Pappas, G. (2013), “Stochastic game approach for replay attack detection,” in *IEEE 52nd Annual Conference on Decision and Control (CDC)*, pp. 1854–1859.
- Miao, F., Zhu, Q., Pajic, M., and Pappas, G. J. (2017), “Coding Schemes for Securing Cyber-Physical Systems Against Stealthy Data Injection Attacks,” *IEEE Trans. on Control of Network Systems*, 4, 106–117.
- Mo, Y. and Sinopoli, B. (2010), “False Data Injection Attacks in Control Systems,” in *First Workshop on Secure Control Systems, CPS Week*.
- Mo, Y. and Sinopoli, B. (2016), “On the performance degradation of cyber-physical systems under stealthy integrity attacks,” *IEEE Transactions on Automatic Control*, 61, 2618–2624.
- Mo, Y., Garone, E., Casavola, A., and Sinopoli, B. (2010), “False data injection attacks against state estimation in wireless sensor networks,” in *49th IEEE Conf. on Decision and Control (CDC)*, pp. 5967–5972.
- Mo, Y., Kim, T.-H., Brancik, K., Dickinson, D., Lee, H., Perrig, A., and Sinopoli, B. (2012), “Cyber-physical security of a smart grid infrastructure,” *Proceedings of the IEEE*, 100, 195–209.
- Mo, Y., Weerakkody, S., and Sinopoli, B. (2015), “Physical authentication of control systems: designing watermarked control inputs to detect counterfeit sensor outputs,” *Control Systems*, 35, 93–109.

- Pajic, M., Weimer, J., Bezzo, N., Tabuada, P., Sokolsky, O., Lee, I., and Pappas, G. (2014), “Robustness of attack-resilient state estimators,” in *ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS)*, pp. 163–174.
- Pajic, M., Lee, I., and Pappas, G. J. (2017), “Attack-Resilient State Estimation for Noisy Dynamical Systems,” *IEEE Transactions on Control of Network Systems*, 4, 82–92.
- Pasqualetti, F., Dorfler, F., and Bullo, F. (2013), “Attack detection and identification in cyber-physical systems,” *IEEE Transactions on Automatic Control*, 58, 2715–2729.
- Ploeg, J., Semsar-Kazerooni, E., Lijster, G., van de Wouw, N., and Nijmeijer, H. (2015a), “Graceful degradation of cooperative adaptive cruise control,” *IEEE Trans. on Int. Trans. Sys.*, 16, 488–497.
- Ploeg, J., Semsar-Kazerooni, E., Lijster, G., van de Wouw, N., and Nijmeijer, H. (2015b), “Graceful Degradation of Cooperative Adaptive Cruise Control,” *IEEE Trans. on Int. Trans. Sys.*, 16, 488–497.
- R, T., Murguia, C., and Ruths, J. (2017), “Tuning Windowed Chi-Squared Detectors for Sensor Attacks,” *CoRR*, abs/1710.02573.
- Rajamani, R. (2011), *Vehicle dynamics and control*, Springer Science & Business Media.
- Shepard, D., Bhatti, J., and Humphreys, T. (2012), “Drone Hack,” *GPS World*, 23, 30–33.
- Shoukry, Y. and Tabuada, P. (2016), “Event-Triggered State Observers for Sparse Sensor Noise/Attacks,” *IEEE Transactions on Automatic Control*, 61, 2079–2091.
- Shoukry, Y., Chong, M., Wakaiki, M., Nuzzo, P., Sangiovanni-Vincentelli, A. L., Seshia, S. A., Hespanha, J. P., and Tabuada, P. (2016), “SMT-Based Observer Design for Cyber-Physical Systems Under Sensor Attacks,” in *Int. Conference on Cyber-Physical Systems (ICCPS)*, pp. 1–10.
- Smith, R. (2011), “A decoupled feedback structure for covertly appropriating networked control systems,” *Proceedings of IFAC World Congress*, pp. 90–95.
- Snider, J. M. et al. (2009), “Automatic steering methods for autonomous automobile path tracking,” *Robotics Institute, Pittsburgh, PA, Tech. Rep. CMU-RITR-09-08*.
- Sundaram, S. (2012), *Fault-tolerant and secure control systems*, University of Waterloo, Lecture Notes.
- Sundaram, S., Pajic, M., Hadjicostis, C., Mangharam, R., and Pappas, G. (2010), “The Wireless Control Network: Monitoring for Malicious Behavior,” in *49th IEEE Conference on Decision and Control (CDC)*, pp. 5979–5984.

- Teixeira, A., Pérez, D., Sandberg, H., and Johansson, K. H. (2012), “Attack models and scenarios for networked control systems,” in *Int. Conf on High Confidence Networked Systems (HiCoNS)*, pp. 55–64.
- Teixeira, A., Sou, K. C., Sandberg, H., and Johansson, K. H. (2015), “Secure control systems: A quantitative risk management approach,” *IEEE Control Systems*, 35, 24–45.
- Tippenhauer, N. O., Pöpper, C., Rasmussen, K. B., and Capkun, S. (2011), “On the requirements for successful GPS spoofing attacks,” in *18th ACM Conf. on Computer and Com. Security, CCS*, pp. 75–86.
- Zetter, K. (2016), “Inside the Cunning, Unprecedented Hack of Ukraine’s Power Grid, *Wired Magazine*,” .