KTH Electrical Engineering

Tuning of Anomaly Detectors in the Presence of Sensor Attacks

DAVID UMSONST

Licentiate Thesis
Stockholm, Sweden, 2019

KTH Royal Institute of Technology
School of Electrical Engineering and Computer Science
Division of Decision and Control Systems
SE-100 44 Stockholm
SWEDEN

TRITA-EECS-AVL-2019:66
ISBN: 978-91-7873-289-0

Akademisk avhandling som med tillstånd av Kungliga Tekniska högskolan framlägges till offentlig granskning för avläggande av teknologie licenciatexamen i reglerteknik fredagen den 11 oktober 2019 klockan 10.00 i sal V2 Kungliga Tekniska högskolan, Teknikringen 76, KTH, Stockholm.

Tryck: Universitetsservice US AB

Abstract

Critical infrastructures, such as the power grid and water distribution networks, are the backbone of our modern society. With the integration of computational devices and communication networks in critical infrastructures, they have become more efficient, but also more vulnerable to cyberattacks. Due to the underlying physical process, these cyberattacks can not only have a financial and ecological impact, but also cost human lives. Several reported cyberattacks on critical infrastructures show that it is vital to protect them from these attacks. Critical infrastructures typically rely on accurate sensor measurements for optimal performance. In this thesis, we, therefore, look into attacks that corrupt the measurements.

The first part of the thesis is concerned with the feasibility of a worst-case sensor attack. The attacker's goal is to maximize its impact, while remaining undetected by an anomaly detector. The investigated worst-case attack strategy needs the exact controller state for its execution. Therefore, we start by looking into the feasibility of estimating the controller state by an attacker that has full model knowledge and access to all sensors. We show that an unstable controller prevents the attacker from estimating the controller state exactly and, therefore, makes the attack non-executable. Since unstable controllers come with their own issues, we propose a defense mechanism based on injecting uncertainty into the controller. Next, we examine the confidentiality of the anomaly detector. With access to the anomaly detector state, the attacker can design a more powerful attack. We show that, in the case of a detector with linear dynamics, the attacker is able to obtain an accurate estimate of the detector's state.

The second part of the thesis is concerned with the performance of anomaly detectors under the investigated attack in the first part. We use a previously proposed metric to compare the performance of a χ^2, cumulative sum (CUSUM), and multivariate exponentially weighted moving average (MEWMA) detectors. This metric depends on the attack impact and average time between false alarms. For two different processes, we observe that the CUSUM and MEWMA detectors, which both have internal dynamics, can mitigate the attack impact more than the static χ^2 detector. Since this metric depends on the attack impact, which is usually hard to determine, we then propose a new metric. The new metric depends on the number of sensors, and the size of an invariant set guaranteeing that the attack remains undetected. The new metric leads to similar results as the previously proposed metric, but is less dependent on the attack modeling. Finally, we formulate a Stackelberg game to tune the anomaly detector thresholds in a cost-optimal manner, where the cost depends on the number of false alarms and the impact an attack would cause.

Sammanfattning

Kritiska infrastrukturer, så som elnätet eller vattenförsörjningssystemet, är ryggraden i vårt moderna samhälle. Effektiviteten av kritiska infrastrukturer har ökats genom integration med beräkningsenheter och kommunikationsnätverk, men detta har medfört att de också har blivit mer sårbara för cyberattacker. På grund av den underliggande fysikaliska processen kan dessa cyberattacker inte bara ha ekonomiska och ekologiska effekter, utan de kan också kosta människoliv. Flera rapporterade cyberattacker mot kritiska infrastrukturer visar att det är viktigt att skydda dem från dessa attacker. Kritiska infrastrukturer förlitar sig vanligtvis på noggranna sensormätningar för optimal prestanda. I denna avhandling undersöker vi därför attacker som korrumperar mätningar.

Den första delen av avhandlingen handlar om genomförandet av en sensorattack i ett värstafallsscenario. Angriparens mål är att maximera verkan av attacken, medan den förblir oupptäckt av en feldetektor. Den undersökta värstafallstrategin behöver exakt information av regulatorns tillstånd för att kunna användas. Därför börjar vi med att titta på möjligheten att en angripare ska kunna uppskatta regulatorns tillstånd samtidigt som den känner till modellen och har tillgång till alla sensorer. Vi visar att en instabil regulator förhindrar angriparen från att exakt uppskatta regulatorns tillstånd och därmed förhindrar attacken. Eftersom instabila regulatorer introducerar andra problem, föreslår vi en försvarsmekanism baserad på injektion av osäkerhet i regulatorn. Därefter undersöker vi feldetektorns konfidentialitet. Med kännedom om feldetektorns tillstånd kan angriparen skapa en kraftfullare attack. Vi visar att angriparen kan få en noggrann uppskattning av detektorns tillstånd när detektorn har linjär dynamik.

Den andra delen av avhandlingen behandlar feldetektorers prestanda medan de utsätts för de attacker som introducerades i första delen. Vi använder en tidigare föreslagen metrik för att jämföra prestandan av detektorer baserade på χ^2-fördelningen, kumulativ summa (CUSUM), och multivariat exponentiellt viktat glidande medelvärde (MEWMA). Denna metrik beror på verkan av attacken och genomsnittlig tid mellan falska larm. Vi observerar att CUSUM- och MEWMA-detektorerna, där båda har intern dynamik, kan begränsa verkan av attacker bättre än vad den statiska χ^2-detektorn kan för två olika processer. Eftersom denna metrik beror på attackens verkan, vilket vanligtvis är svårt att fastställa, föreslår vi en ny metrik. Den nya metriken beror på antalet sensorer och storleken på en invariant mängd som garanterar att attacken förblir oupptäckt. Den nya metriken leder till liknande resultat som den tidigare föreslagna metriken, men är mindre beroende av en modell av angriparen. Slutligen formulerar vi ett Stackelberg-spel för att ställa in trösklar för feldetektorn på ett kostnadsoptimalt sätt, där kostnaden beror på antalet falska larm och potentiell verkan av attacker.

Sometimes science is more art than science, Morty.
- Rick Sanchez

Acknowledgments

First and foremost, I would like to thank my supervisors: Henrik Sandberg, for his guidance to focus my path continuously when wandering aimlessly through the fields of research, and Karl-Henrik Johansson, for radiating positivity.

I am also very grateful for all my amazing, former and present, colleagues at our department, with whom I can share both the joy and the suffering that come with working in academia! Thanks to Jezdimir Milošević for proofreading my thesis and making me conversational in Serbian, Mladen Čičić for sharing his enthusiasm about craft beer with me, Erik Berglund for being a boardgame wizard, Lars Lindemann for our walks, talks, and good times in our conference hotel rooms, and Rui Oliveira for being a constant in our changing office environment. Next I would like to thank Emma Tegling for her advice and help, Sebin Gracy for proofreading my thesis, Michelle Chong for culinary experiences and proofreading my thesis, and Dirk van Dooren for making the tedious task of correcting homework enjoyable. Also thanks to Andrea Bisoffi, Alexander Johansson, and Ehsan Nekouei for the trip to Chicago, and Joana Fonseca, Mina Ferizbegović, and Pian Yu for helping me to keep my office plants alive. Thanks to Matias Müller, Rodrigo Gonzalez, and Pedro Roque for the jam sessions, Joakim Björk for the climbing and midsummer adventures, Rijad Ališić and Ingvar Ziemann for being great colleagues and proofreading parts of my thesis, and, respectively, helping me translate my abstract into Swedish and helping me with some of the theoretical proofs. Thanks to Elis Stefansson for the fun lunches, Christos Verginis for making commuting more pleasant, Valerio Turri for the fun times spent on his rooftop and Inês Lourenço for sharing her optimism, enthusiasm, and joy with me.

Moreover, I want to thank my former corridor mates, Christoforos and Vangelis, for being good friends and making Stockholm more enjoyable.

Further, I want to express my gratitude to my friends in my hometown, especially the members of my 'bowling' team KC Die Flummis, and also Christin, Maikel, and Sarah, who always welcomed me back with open arms and made me feel like I have never left. These episodes of normality back home really helped me through the rocky parts of my PhD life.

Finally, I want to thank my family for their love and support during my whole life.

David Umsonst

Contents

Notation

$\mathbb{R}$	Set of real numbers
$\mathbb{R}_{\geq a}$	Set of real numbers greater than or equal to $a \in \mathbb{R}$
$\mathbb{R}^n$	Set of real n–dimensional vectors
$\mathbb{R}^{n \times m}$	Set of real $(n \times m)$–dimensional matrices
I_n	$(n \times n)$–dimensional identity matrix
0	The number zero, a zero vector, or a zero matrix
$A \geq 0$	Positive semi-definite matrix A
$A > 0$	Positive definite matrix A
$\rho(A)$	Spectral radius of matrix A
$\sigma_{\max}(A)$	Maximum singular value of matrix A
$A^\dagger$	Moore-Penrose pseudo inverse of matrix A
$\|\|x\|\|_2$	Euclidean norm of a vector x
$\|\|x\|\|_\infty$	Infinity norm of a vector x
$\lceil x \rceil$	Rounds a real number x up to the next integer
$\mathcal{N}(\mu, \Sigma)$	Gaussian distribution with mean μ and covariance Σ
$\mathbb{E}\{x\}$	Expected value of a random variable x
$\mathrm{Var}\{x\}$	Variance of a random variable x
$\mathrm{supp}(f)$	Support of a real-valued function f

Chapter 1

Introduction

1.1 Motivation

Humans have never been healthier, been more educated, and lived longer (see Figure 1.1). One reason for this improvement of the human living condition is that, to cover their basic needs, humans have developed critical infrastructures including power grids, road networks, and water distribution networks [2]. The introduction of computational devices and communication networks further improved the way the critical infrastructures work and made the coverage of basic needs more efficient, cheaper, and cleaner. The smart grid, intelligent transportation systems, and smart water distribution networks are just some of the examples of advanced critical infrastructures in the near future. Since critical infrastructures consist of physical

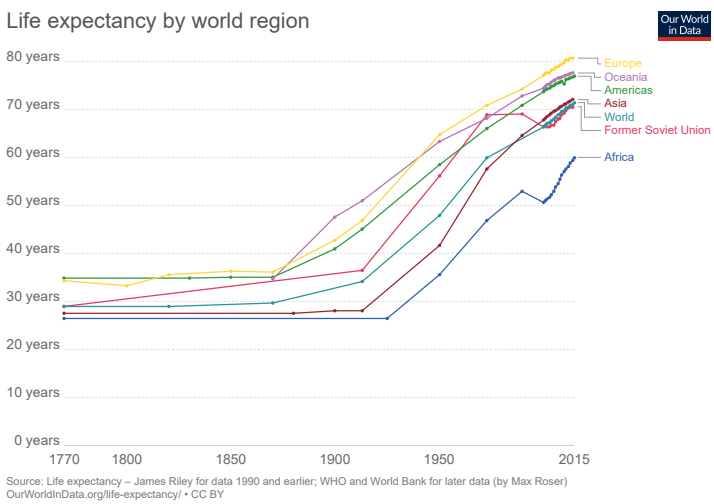

Figure 1.1: Positive trend of life expectancy [1].

processes, computational devices and communication networks, we call them cyber-physical systems (CPSs). However, due to the connection of the processes via communication networks, CPSs become vulnerable to cyberattacks. Attackers can exploit vulnerabilities in the cyber space and reach havoc on a physical level, if the CPSs are not properly protected.

At a first glance, cyberattacks on our critical infrastructures may seem like a scenario that would happen in a dystopian future. Unfortunately, these attacks are already happening. Hemsley *et al.* [3] provide a list of conducted cyberattacks, where the earliest attack dates back to 1903. Two attacks that were aimed directly at critical infrastructures are the attack on the Maroochy water services [4, 5] and the attack on the Ukrainian power grid [6, 7]. The attack on the Maroochy water services caused the spillage of nearly a million liters of untreated water into a storm drainage and damaged the marine life, while the attack on the Ukrainian power grid left more than 200000 customers without electricity. Since the integration of computational devices in our industrial processes is progressing as well, not only critical infrastructures but also industrial processes are endangered. The Stuxnet worm is a prominent example of attacks on industrial processes and was designed to target an Iranian uranium enrichment facility [8]. Other notable incidents that made it into the popular media are hackers that remotely activated the brakes of a jeep [9], a demonstration of researchers from Tulsa University on how easily it was to hack wind farms [10], and the derailing of local trams due to a hack by a teenager [11].

Two, if not the most dangerous, aspects of cyberattacks are that coordinated attacks are much cheaper to conduct than physical attacks and the attacker does not have to be physically present to conduct the attack. The perpetrator basically just needs a computer to attack a system that might be located at the other side of the world. Therefore, the adversary model presented in [12] does not only include terrorist groups and nation states, but also disgruntled employees as in the case of the Maroochy attack [5]. Even teenagers could launch a cyberattack without being aware of the damage they might cause, as in the case of the tram hack [11].

Therefore, it is of utmost importance to secure the critical infrastructure and industrial processes. Several governments have published strategies to protect critical infrastructures, such as Sweden [13], the United States of America [14], and Germany [2]. Security measures based on information technology (IT), such as authentication and encryption, are one way to secure industrial processes. However, and due to real time requirements and legacy equipment, utilizing these IT measures is not always feasible for CPS [15]. Further, these IT measures alone will not block all possible attack vectors

In this thesis, we, therefore, utilize a physical model of a CPS to provide an additional layer of protection. One key feature of CPS is their closed-loop operation (see Figure 1.2). Sensors measure important plant quantities and send them over a network to a controller or control center, where, based upon these measurements, actuator signals are determined to control the plant in an optimal or efficient manner. Since faults, like actuator or sensor faults, can happen randomly, a CPS is

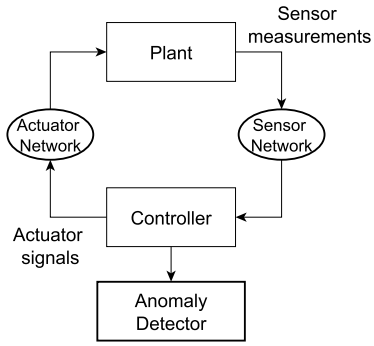

Figure 1.2: Closed-loop behaviour of a CPS

usually equipped with an anomaly detector [16]. The anomaly detector analyses the measurements and the actuator signals based on a model of the plant to see if a fault has occurred. While faults are random in nature, attacks are maliciously designed and want to avoid detection. In the last decade, the field of cyber-physical security has investigated these attacks and defenses against them. For CPSs the content of the data sent over a network has to obey the physical laws governing the plant. For example, a temperature measurement may not change 100 °C in a second. Hence, these cyber-physical security investigations often use the physics of the plant to determine if someone has changed the measurement or actuator signals, or estimate the possible impact of an attacker who wants to remain undetected and we call attackers that try to remain undetected *stealthy*. These new types of security measures should be seen as a complement to the IT security measures rather than a replacement.

1.2 Research Challenges

In this thesis, we focus on sensor attacks on control systems equipped with an anomaly detector (see Figure 1.3). In the following, we will present several research challenges (**C**) when it comes to sensor attacks and describe what exactly we mean by sensor attacks.

The attacker that has access to the measurements can add an additive signal $y_a(k)$ to the measurement signal $y(k)$ such that $\tilde{y}(k) = y(k) + y_a(k)$. For example, by reducing the value of a pressure measurement, the attacker deceives the controller into increasing the pressure of a tank. This might result in an explosion in the plant. Figure 1.3 shows us that the operator uses an anomaly detector, though. Hence, $y_a(k)$ can not be arbitrarily designed, if the attacker wants to avoid detection as well. If the operator does not detect the attacker, it cannot employ countermeasures and mitigate the attack impact. Hence, the desire to remain undetected is often assumed for an attacker.

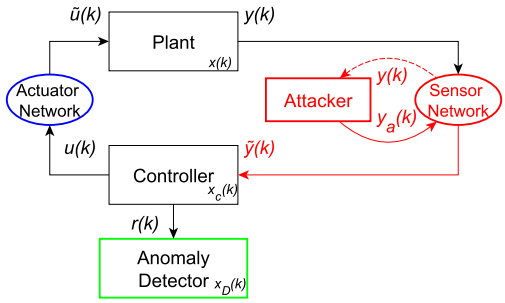

Figure 1.3: Block diagram of the closed-loop system equipped with an anomaly detector under a sensor attack.

Sensor attacks have attracted considerable attention [17–24]. For example, Mo *et al.* [22] define a notion of a perfectly attackable system and present conditions for when a plant is perfectly attackable. Sensor attacks that maximize the error covariance matrix of a state estimator are investigated in [23], while Cárdenas *et al.* [24] investigate three different sensor attacks under two different anomaly detector policies.

It is often assumed that the attacker has some knowledge about the dynamical model of the closed-loop system. Furthermore, the attacker is often assumed to have additional knowledge such as the whole system state [22], the internal state of the controller $x_c(k)$, [19, 23], the detector's state $x_D(k)$, [19], or some internal estimate of the output [24], when the attack starts. This knowledge helps the attacker to remain undetected.

While knowledge about the plant, controller, and detector model could be obtained by acquired documentation of the closed-loop system, the internal states cannot be available when the attack starts. The state of the controller and detector at the start of the attack depend on the whole system's past inputs, which are unknown to the attacker. Therefore, the knowledge of $x_c(k)$ and $x_D(k)$ is impossible for the attacker to have when it has just gained access to the measurements. This leads us to the first research challenge.

C1: Is it possible for the attacker to obtain knowledge the internal state, e.g. $x_c(k)$ and $x_D(k)$, after attacker has gained access to the measurements when the closed-loop system and detector dynamics are known?

Next, assuming the attacker is able to execute its stealthy attack, it is crucial to know how large the impact of a sensor attack is and how to mitigate it. Hence, the second challenge we address is related to impact estimation.

C2: What is the worst-case attack impact of a stealthy sensor attack?

In [17, 18] the impact of the attack has can be interpreted as the volume of the set of reachable states by the attacker. The reachable set is then overapproximated

for the case of a χ^2 detector. However, the worst-case impact will be influenced by the detector, because the attacker designs $y_a(k)$ to remain undetected. Hence, the choice of the detector is an important design criteria for the operator. Therefore, the third challenge concerns the comparison of anomaly detectors.

C3: What is a good metric to compare the performance of anomaly detectors under attack? Which detector mitigates the attack impact the most?

The metric for detector comparison proposed in [25] is one way to compare the performance of detectors, but to the best of our knowledge it has not yet been used to compare other detectors than the χ^2 and cumulative sum (CUSUM) detector. Hence, the comparison of other detectors using this metric is still open. However, this metric depends on the attack impact, which might differ depending on the assumptions made for the attacker. Therefore, it would be good to have a metric for detector comparison that does not depend on the attack impact.

Furthermore, attacks are not occurring permanently and the operator has to consider the normal working conditions of the plant as well. Therefore, it is not only important to choose the detector that mitigates the attack impact the most, but also to tune it to reduce costs during nominal operation, for example, the cost induced by false alarms. Although Ghafouri *et al.* [26] present a Stackelberg game for the choice of a cost optimal detector tuning, this game does not fit our sensor attack framework with stealthy attacks. This leads us to the final challenge we address.

C4: What is the optimal detector tuning for the operator to avoid high operational costs in the case of stealthy sensor attacks?

Throughout the course of this thesis we will tackle these research challenges.

1.3 Thesis Outline and Contributions

In this section, we present the structure of the thesis and provide summaries of each chapter.

Chapter 2: Literature Overview

In this chapter, we give an overview of not only the literature on sensor attacks but over the whole field of CPS security. We begin by looking into different attack angles and then look at proposed the defense mechanisms. The defense mechanisms reviewed include adjusting controllers, resource allocation, combining IT security measures with control system, and game-theoretic approaches. Last, the problem of privacy in control systems is reviewed.

Chapter 3: A Modeling Framework for Sensor Attacks

In this chapter, we introduce the framework for the sensor attack scenario, which is used throughout the thesis. We start by defining a discrete-time linear plant and controller. Although the detector model we propose is general, we also provide the definition of three specific detectors, namely the χ^2, multivariate exponentially moving average (MEWMA), and CUSUM detectors. Then, assumptions on the attacker's knowledge, the attack strategy, as well as the definition of the worst-case attack impact are stated. The chapter ends with presenting the dynamics of a closed-loop system with an observer-based controller under the sensor attack, since this controller is used in all the illustrative examples.

Chapter 4: On the Confidentiality of the Controller State

Chapter 4 tackles **C1** by investigating when the attacker is able to perfectly estimate the controller's internal state $x_c(k)$. We show that a necessary and sufficient condition for the attacker to perfectly estimate $x_c(k)$ is that the controller has no eigenvalues outside of the unit circle. Furthermore, we specify all observer gains for a non-optimal observer that perfectly estimates $x_c(k)$ when all controller eigenvalues are inside the unit circle. We discuss how adding noise to the controller input prevents the attacker from obtaining a perfect estimate. Further, we argue that using an unstable controller as a defense against this confidentiality attack is only a good idea for specific plants. An illustrative example with a three tank process verifies the results of this chapter.

The chapter is based on the publication:

- D. Umsonst, H. Sandberg, "On the confidentiality of controller states under sensor attacks," *Under journal review, 2019*

Chapter 5: On the Confidentiality of the Detector State

Challenge **C1** is also tackled in Chapter 5, although the focus lies here on the confidentiality of the detector state $x_D(k)$. The analysis is limited to detectors with linear dynamics. We show that an attacker is able to estimate $x_D(k)$ and we provide a quality bound for the estimate. Furthermore, while estimating the detector state the attacker is able to inject an attack signal that mimics the statistics of the detector output and simultaneously has an impact on the plant. Mimicking the statistics will not raise the suspicion of an operator that is watching the detector output in the control center. A benchmark model for the excitation of tall buildings by wind with a MEWMA detector is used to verify the results of this chapter.

The chapter is based on the publication:

- D. Umsonst, E. Nekouei, A. Teixeira, H. Sandberg, "On the confidentiality of linear anomaly detector states," *in Proceedings of the American Control Conference (ACC), 2019.*

Chapter 6: Comparison of Detectors

In this chapter, we tackle **C2** and **C3** and the chapter is split into two parts. In the first part, we compare the χ^2, CUSUM, and MEWMA detectors using the metric proposed in [25]. We first show how the impact of a stealthy attack can be determined for these three detectors and then plot the metric for the three tank process used in Chapter 4 and for a quadruple tank process. We observe that the CUSUM and MEWMA detectors perform better than the χ^2 detector, i.e. mitigate the attack impact more. However, the tuning of the CUSUM detector is crucial to its performance. In the second part of the chapter, we propose a new metric for the comparison of detectors under sensor attacks that does not depend on the plant dynamics or the attacker's objective. This new metric is used to compare the three detectors again and the metric yields similar results to the previous detector comparison in this chapter.

The chapter is based on the publications:

- D. Umsonst, H. Sandberg, A. A. Cárdenas, "Security analysis of control system anomaly detectors," *in Proceedings of the American Control Conference (ACC), 2017.*
- D. Umsonst, H. Sandberg, "Anomaly detector metrics for sensor data attacks in control systems," *in Proceedings of the American Control Conference (ACC), 2018.*
- D. Umsonst, H. Sandberg, "A game-theoretic approach for choosing a detector tuning under stealthy sensor data attacks," *in Proceedings of the 57th Conference on Decision and Control (CDC), 2018.*

Chapter 7: A Game-Theoretic Approach to Detector Tuning

The last technical chapter tackles **C4**. While the previous chapter compared the performance of detectors, this chapter considers the optimal tuning of the detector. A Stackelberg game is used to determine the optimal detector threshold. In this game, the defender plays first by choosing a threshold for the detector, while the attacker follows the move by attacking the system. The threshold is chosen to minimize a cost which depends on the number of false alarms and the attack impact. The existence of a solution to the Stackelberg games is shown and conditions for the uniqueness of the solution are presented. We verify the framework for a χ^2 detector and use the quadruple tank process as an application example.

The chapter is based on the publication:

- D. Umsonst, H. Sandberg, "A game-theoretic approach for choosing a detector tuning under stealthy sensor data attacks," *in Proceedings of the 57th Conference on Decision and Control (CDC), 2018.*

Chapter 8: Conclusions and Future Work

In this chapter, we conclude the thesis by summarizing the results and state possible directions for future work.

Author Contributions and Other Publications

In the aforementioned peer-reviewed articles, the author of the thesis has formulated and solved most of the problems, and written the papers. The results of the coauthors are clearly indicated throughout the thesis.

The following publications in which the author of the thesis participated are not covered in the thesis:

- F. Kintzler, T. Gawron-Deutsch, S. Cejka, J. Schulte, M. Uslar, E. MSP Veith, E. Piatkowska, P. Smith, F. Kupzog, H. Sandberg, M. S. Chong, D. Umsonst, M. Mittelsdorf, "Large Scale Rollout of Smart Grid Services," *2018 Global Internet of Things Summit (GIoTS), 2018.*

- J. Milošević, D. Umsonst, H. Sandberg, K. H. Johansson "Quantifying the Impact of Cyber-Attack Strategies for Control Systems Equipped With an Anomaly Detector," *in Proceedings of the European Control Conference (ECC), 2018.*

- M. S. Chong, D. Umsonst, H. Sandberg, "Voltage regulation of a power distribution network in a radial configuration with a class of sector-bounded droop controllers," *in Proceedings of the 58th Conference on Decision and Control (CDC), 2019. (accepted)*

- M. S. Chong, D. Umsonst, H. Sandberg, "Local Voltage Control of an Inverter-Based Power Distribution Network with a Class of Slope-Restricted Droop Controllers," *in Proceedings of the 8th IFAC Workshop on Distributed Estimation and Control in Networked Systems (NecSys), 2019.*

- B. Kang, D. Umsonst, M. Faschang, C. Seitl, I. Friedberg, F. Kupzog, H. Sandberg, K. McLaughlin, "Intrusion Resilience For PV Inverters In a Distribution Grid Use-case Featuring Dynamic Voltage Control," *in Proceedings of the 14th International Conference on Critical Information Infrastructures Security (CRITIS) , 2019.*

Remark 1.1. The work with Milošević *et al.* extends the impact estimation results of Chapter 6 to a broader class of attacks and detectors.

Chapter 2

Literature Overview

In this chapter, we give an overview of the existing literature of security of cyber-physical systems. Lun *et al.* [27] show in their survey paper how the number of publications in the field of security for cyber-physical systems is rapidly increasing. Therefore, this is by no means a comprehensive overview.

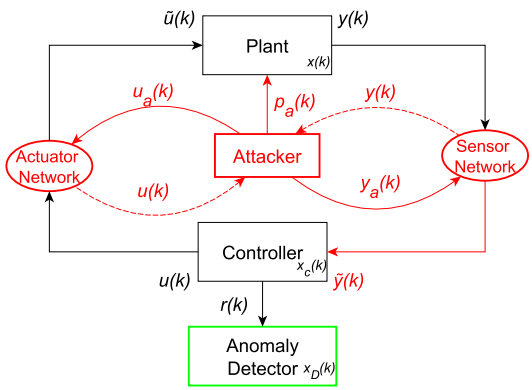

Figure 2.1: Block diagram of the closed-loop system equipped with an anomaly detector under an attack.

Further, in this chapter, we will not only consider sensor attacks, but also look into other attack strategies. Figure 2.1 shows a closed-loop system, where the attacker has more resources than for a sensor attack. The attacker is able to listen to the actuator signals, change them with $u_a(k)$, and to launch physical attacks $p_a(k)$.

2.1 Attack Vectors and Attack Strategies

Before we are able to design defenses against attacks on control systems, we need to know what threats we are facing. Further, understanding possible attacks is one of the challenges of CPS security according to [12]. Therefore, a lot of research is conducted on possible attack vectors and attack strategies.

Teixeira *et al.* [28] propose an attack space that is spanned by the attacker's disclosure and disruptive resources, and its model knowledge. Disruptive resource, $y_a(k)$ and $u_a(k)$, are used to change measurements and actuator signals, respectively. Further, the attacker can cause some physical damage ($p_a(k)$) to the plant, while disclosure resources (dashed lines in Figure 2.1) are used to listen in to the measurements and actuator signals and infer some additional knowledge about the process. The model knowledge about the closed-loop system helps the attacker to launch more sophisticated attacks. While [28] proposes a framework to model attacks which uses linear discrete-time system, Pasqualetti *et al.* [29] use continuous-time descriptor systems in their modeling of attacks. Further, conditions for the detectability and identifiability of attacks are given in [29]. Duz *et al.* [30] extend the conditions for detectability to linear-impulsive systems, i.e. linear systems that include jumps in their dynamics. The frameworks in both [28] and [29] are able to cover several attack strategies. Therefore, we will look at some attack strategies that have been investigated in the following.

Our overview of attack strategies begins with the *covert attack* presented by Smith [31]. This attack is very powerful, because the attacker has access to all measurements and actuators and full model knowledge. In [31], a feedback structure for the covert attack is presented, where the attacker is able to remove the changes induced at the plant from the measurements. In that way, the attack is not visible at the control center, which makes the attack very dangerous. Another attack strategy that is not visible at the output of the plant is the *zero-dynamic attack* [28, 29]. In a nutshell, zero dynamics are plant internal dynamics, which can be excited with a nonzero input and result in a zero output. In case the zero dynamics have unstable modes, an attacker is able to excite these unstable modes without a changing the plant's output. At a first glance, an attacker needs to know the exact model of the plant to excite the zero dynamics with its attack. But [28] presents the possibility of local zero dynamic attacks, which only need partial model knowledge, and Park *et al.* [32] show how an attacker can employ robust control techniques to launch a zero-dynamic attack with uncertain model knowledge. In optimal control, the controller input is typically designed to minimize a certain optimization criterion, for example the fuel consumption of a plane or the time it takes to fulfill a task. Lipp *et al.* [33] design attacks by maximizing the criteria that are typically minimized in control applications and they call these attacks *antagonistic control*. One particular example of an attack is called *ambush control*. Here, the attacker remains undetected for a certain amount of time to set the stage for the actual attack. Then the attacker tries to maximize the criterion without the concern of remaining undetected. *False-data injection attacks* are attacks where the

attacker inject artificial data into the system by changing the actuator and/or sensor measurements. Many of the attacks mentioned before can be seen as false-data injection attacks, for example the ambush control of [33] and the zero-dynamics attacks in [28]. Other false-data injection attacks are investigated in [19, 22, 23].

An attack without the need of model knowledge are *replay attacks* considered, for example, in [28, 34]. Here, the attacker records sensor measurement during a normal run of the control system. Then these measurements are replayed while the attacker changes the actuator signals. Therefore, the measurements, which the controller receives, appear normal, while the plant might behave in a undesired way. Replay attacks have already been conducted in the real world, since the Stuxnet worm used an attack strategy similar to the replay attack in its execution [8]. Another attack, which does not require model knowledge, is a *Denial-of-Service (DoS) attack* as, for example, investigated in [35–37]. As the name suggests the attacker blocks the communication between the controller and the plant in this attack strategy. This attack is especially severe if the open-loop system is unstable, because then the blocking of for example measurements leads to an open-loop configuration. In the case of event triggered control, DoS attacks are also fatal because there communication is only established when required. The lack of required communication due to the DoS attack in these cases might also lead to instability. The DoS attack can be modeled as completely random [35], as an attacker determining the package drop probability of the network [36], or as a sequence of intervals in which communication is not possible [37].

The attack strategies presented up to now targeted either the integrity or the availability of the control system data. Another attack angle is to target the confidentiality of the control system data. Xue *et al.* [38] investigate the confidentiality of the internal state of a double integrator network, which can be interpreted as a network of autonomous vehicles. Yuan *et al.* [39] look into an attacker that tries to identify the controller structure, while in [40] the adversary tries to identify the controller's gain in the setting of a wide area power system.

2.2 Attack Mitigation

After giving an overview over several attack strategies, let us now look at possible defense mechanisms. A natural way to start is to investigate how the already existing infrastructure, such as controllers and detectors, can be used to detect an attack or mitigate its impact. Kafash *et al.* [17] use an artificial actuator saturation to limit the set of reachable states of the attacker, while [18] adjusts the controller to limit the reachable set of a stealthy attacker. In [39], an appropriate controller design is proposed to prevent an attacker from obtaining the controller structure. The anomaly detector of the control system is of interest in the attack mitigation and detection as well. Other than randomly occurring faults, such as sensor and actuator failures, attacks are intelligently designed and might avoid detection and simultaneously cause the worst possible impact. Therefore, it is of interest to

investigate which detector should be used to mitigate the attack impact the most. Urbina *et al.* [25] propose a metric to compare different detectors by plotting the impact of a stealthy attack over the mean time between false alarms. Murguia *et al.* [19] compare the attack impact under a χ^2 detector and a CUSUM detector and it turns out that the CUSUM detector in the right configuration mitigates the attack impact more than the χ^2 detector.

It is important for the operator to have a good knowledge of what is going on in the plant, therefore state estimators can be used to monitor the internal variables of the plant. However, if an attacker changes the measurements this state estimate might diverge from the actual state. Therefore, secure state estimators have been investigated in [41, 42]. Both [41] and [42] show that for discrete-time and continuous-time systems, respectively, a fundamental limit to recover the state from an attack is that the attacker cannot attack more than half of the sensors in the system.

Hence, [41] and [42] show us that an operator needs to protect at least half of its sensors to guarantee a secure state estimation. To do so an operator has to allocate security measures to protect the sensors. Security indices shows us how vulnerable a sensor [43], actuator [44], or even the whole system [45, 46] is and, therefore, give an indication on where to allocate security measures. However, an operator might have limited resources and needs to choose which of the sensors to protect. Milošević *et al.* [47] provide a framework for security measure allocation that minimizes the allocation cost of the operator.

2.3 Defense Mechanisms with Roots in IT Security

In this section, we will discuss defense mechanisms that can also be found in IT security application. Some of these techniques are called active defense mechanisms. Active means that the system or signals are actively modified to, for example, detect the attacker or mitigate the attack impact. *Authentication* can be seen as an active defense mechanism, because an authentication message or signal is artificially added to the signals to ensure their integrity. In IT security, it is common to add a message authentication code to each message to ensure its integrity. However, due to computational limitations and real time requirements, we might not be able to use data authentication for each data sample. A remedy for that is to use irregular authentication of the data signals as proposed in [48, 49]. This way the real time requirements can still be fulfilled, while providing guarantees for the operator's estimation error.

An alternative to adding a message authentication code to each message is to modify the signal itself to provide data integrity. One way to do this is *watermarking*, which is also often used in IT security. It is for example included in a roadmap for software engineering security in the beginning of this century [50]. To the best of our knowledge, two approaches for watermarking have been proposed in the field of CPS security. *Additive watermarking* [20, 21, 34] can be seen as a physical

authentication code. The idea is to add noisy signal with time-varying statistics to the actuator signal such that the statistics of the sensor measurements change. By a malicious change of the measurements this watermarking signal gets removed or changed, which leads to a detection of the attack. The additive watermarking is used in [34] to detect replay attacks, but [20] and [21] show that a more general class of attacks can be detected via watermarking. However, by adding an artificial noise to the actuator signal, the performance of the closed-loop system might degrade. *Multiplicative watermarking* [51–53] is an alternative to the additive approach that does not degrade the system performance. Here, a time-varying watermarking generator is used to give the measurements a watermark on the plant side of the loop and the watermark is removed by a time-varying equalizer on the controller side of the loop. The watermark generator and equalizer are time-varying filters such that an attacker does not know which generator is currently used.

Watermarking can be seen as an artificial uncertainty of the system, which makes it harder for the attacker to stay undetected. *Moving target defense* works along similar lines and tries to increase the attacker's uncertainty about the system model. In power systems, the operator could actively change the topology of the grid to increase the security of the state estimation [54]. This leads to a change in the measurement matrix. Giraldo *et al.* [55] also propose to randomly change the measurements used for the control of a feedback loop. They prove the stability of the closed-loop system and show that this moving target defense leads to the detection of an otherwise stealthy sensor attack.

Another way to use IT security in a control context is to use *homomorphic encryption* as it is, for example, done in [56, 57]. Homomorphic encryption enables an operator to execute the whole feedback loop on encyrpted signals. More specifically, the controller operates on encrypted signals and is encrypted itself. Therefore, only the plant and the operator need to have access to the key for decryption. Farokhi *et al.* [57] show that the closed-loop with homomorphic encryption is stable and that homomorphic encryption can be used to increase the security in a network of agents as well.

2.4 Game Theory and Security

Game theory deals with optimal actions of rational decision makers with not necessarily aligned objectives. Therefore, the scenario of an attacker and defender fits well in the framework of game theory. Game theory has already been applied to the security of networks [58] and also to increase the security of real world environments like airports [59]. Therefore, it is no surprise that game theory has found several applications in the security of cyber-physical systems as, for example, Etesami *et al.* [60] show for dynamic games.

Zhu *et al.* [61] present a cross-layered game-theoretic framework for both the cyber and physical layer of a CPS. This framework is used to design robust control and cybersecurity strategies with resilience in mind. Since additive watermarking

can degrade the performance of the control system, Miao *et al.* [62] provide a game-theoretic approach for the detection of replay attacks with additive watermarking that considers the trade-off between performance and detection. The problem of security allocation can also be posed as a game. Shukla *et al.* [63] present a resource planning game where an attacker tries to destroy communication equipment to create sparsity in the controller and create a loss of performance. A mixed strategy Nash equilibrium to choose the protected equipment is presented to minimize the loss of control performance.

Stackelberg games are a popular game form when it comes to security. A famous application of a Stackelberg game in the security context is the previously mentioned airport security application of game theory [59]. These games are played in two rounds. For a two player Stackelberg game, the leader plays in the first round, while the follower plays in the second round. Since the follower observes the leader's action, the leader has to choose its action such that it minimizes the leader's cost for all possible actions of the follower. Sayin *et al.* [64] use a Stackelberg game for secure sensor design, where it is uncertain if the controller is corrupted or not. Chen *et al.* [65] look into the problem of parameter estimation in the presence of an adversarial sensor and formulate a Stackelberg game to obtain an unbiased estimator with minimum variance. Another Stackelberg game approach can be found in [66], where it is used to maximize the probability to fulfill certain temporal logic constraints under an attack. Ghafouri *et al.* [26] propose a Stackelberg game to decide on a cost-effective detector threshold, which minimizes the cost induced by false alarms and the cost of the attack impact.

2.5 Privacy in Control Systems

In this last part of the literature review, we will not focus on security of CPS but rather their privacy. Privacy can mean not only that an operator wants to keep its information private from an attacker, but also from a curious party. The latter point could be interpreted that a customer wants to provide measurements to its electricity provider without the provider being able to obtain private information about the customer. This problem can be cast as a state estimation problem, such that a privacy preserving mechanism is designed to keep a curious party from estimating private information from the measurements. One way to preserve the privacy is to use coding schemes to encode the measurements sent and decode them at the controller side. Tsiamis *et al.* provide coding schemes based on linear time-varying transformations for stable [67] and unstable systems [68]. They show that in both cases the eavesdropper needs to only miss one of the transmitted to have either the same estimation error as with an open-loop estimation (for a stable system) or an unbounded estimation error (for an unstable system).

Another way to try to preserve privacy is by adding noise to the measurements sent. In that way a single agent can maintain its privacy. However, when adding this additional noise one has to consider the trade-off between the privacy and the

system performance [69]. Furthermore, there is a trade-off between privacy and security as Giraldo *et al.* [70] point out. In [70], it is shown that an attacker can hide its action in the additional noise term. Therefore, the privacy mechanism makes the system less secure.

Information-theoretic concepts like entropy have also become a topic of interest in privacy of control systems as this recent survey shows [71]. Nekouei *et al.* [72], for example, the problem of estimating a common variable from several sensor measurements, where each sensor has a private variable. Two estimation schemes, a local and a global scheme, are discussed and analyzed with respect to a privacy measure based on the conditional entropy.

Chapter 3

A Modeling Framework for Sensor Attacks

In this chapter, we set up a framework that is used to analyze the sensor attacks throughout the thesis. We begin by introducing the plant and controller model, which we use in our investigation of sensor attacks and detector tuning. Then, we propose a general detector model and fit three commonly used anomaly detectors into that model. Finally, the attack model is introduced and we present both the worst-case attack strategy and the worst-case attack impact. Figure 3.1 shows the model setup.

3.1 Plant and Controller Model

Due to the discrete nature of computational devices and networks, we model the plant as a linear discrete-time system

$$\begin{aligned} x(k+1) &= Ax(k) + B\tilde{u}(k) + w(k), \\ y(k) &= Cx(k) + v(k), \end{aligned} \tag{3.1}$$

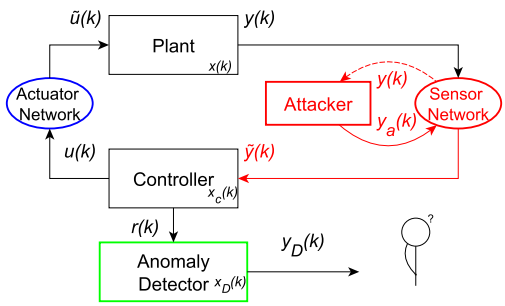

Figure 3.1: Block diagram of the closed-loop system equipped with an anomaly detector under a sensor attack.

where $x(k)$ is the state of the plant in $\mathbb{R}^{n_x}$, $\tilde{u}(k)$ is the actuator signal in $\mathbb{R}^{n_u}$ received at the plant, and $y(k)$ is the measured output in $\mathbb{R}^{n_y}$. Furthermore, $A \in \mathbb{R}^{n_x \times n_x}$ is the system matrix, $B \in \mathbb{R}^{n_x \times n_u}$ is the input matrix, and $C \in \mathbb{R}^{n_y \times n_x}$ is the output matrix. Here, $w(k) \sim \mathcal{N}(0, \Sigma_w)$ is the process noise and $v(k) \sim \mathcal{N}(0, \Sigma_v)$ is the measurement noise, where $\Sigma_w \geq 0$ and $\Sigma_v > 0$ are the covariance matrices of the respective noise term with appropriate dimensions. The noise processes $w(k)$ and $v(k)$ are each independent and mutually uncorrelated. The operator uses an output-feedback controller of the form

$$\begin{aligned} x_c(k+1) &= A_c x_c(k) + B_c \tilde{y}(k), \\ u(k) &= C_c x_c(k) + D_c \tilde{y}(k), \end{aligned} \tag{3.2}$$

where $x_c(k)$ is the controller's state in $\mathbb{R}^{n_c}$, $\tilde{y}(k)$ are the measurements received from the plant, and $u(k)$ is the actuator signal. Here, $A_c \in \mathbb{R}^{n_c \times n_c}$ is the system matrix of the controller, $B_c \in \mathbb{R}^{n_c \times n_y}$ is the input matrix of the controller, $C_c \in \mathbb{R}^{n_u \times n_c}$ is the output matrix of the controller, and $D_c \in \mathbb{R}^{n_u \times n_y}$ is the feedthrough matrix from the received measurements to the actuator signal. The structure (3.2) represents many commonly used controllers.

Note that the measurement and actuator signals sent from the plant and controller, respectively, are not necessarily the signals received by the controller and plant, i.e. $\tilde{y}(k) = y(k)$ and $\tilde{u}(k) = u(k)$ are not necessarily true. This is due to the fact that, for example, packet drops might happen in case the signals are transmitted over a network, an actuator fails or an attacker injects a signal through the network.

Because the operator uses an anomaly detector, as indicated in Figure 3.1, we need a way to determine the detector input $r(k)$. Typically, an estimate of the plant's state is used to determine $r(k)$. Therefore, we assume the following.

Assumption 3.1. The controller state $x_c(k)$ contains an estimate $\hat{x}(k)$ of the plant's state $x(k)$, i.e. $\hat{x}(k) = T_c x_c(k)$, where $T_c \in \mathbb{R}^{n_x \times n_c}$ extracts the estimate from $x_c(k)$.

Note that, for example, in [28] the anomaly detector has its own residual generator, while we incorporated the residual generator in the controller. This means even if the operator is using a PID controller, it needs to have a state estimator in the controller for the anomaly detector.

One controller that fits both (3.2) and Assumption 3.1 is the observer-based controller, with $A_c = A - BK - LC$, $B_c = L$, $C_c = -K$, and $D_c = 0$. In this control strategy $x_c(k) = \hat{x}(k)$, and K and L represent the controller and observer gain, respectively. The observer-based controller is, for example, used in [19] and we will look at it in more detail in Section 3.4.

Assumption 3.2. The plant (3.1) and controller (3.2) are such that

1. (A, B) is stabilizable,
2. (C, A) is detectable,

3. $(A, \Sigma_w^{\frac{1}{2}})$ has no uncontrollable modes on the unit circle, and
4. the controller (A_c, B_c, C_c, D_c) is minimal.

The stability of the closed-loop system depends on the controller matrices A_c, B_c, C_c, and D_c. Therefore, we need the first two points of Assumption 3.2, such that the operator is able to observe and control all unstable modes in the system. The third point is needed in Chapter 4, for example, for the proof of Proposition 4.2. To avoid unnecessary dynamics, the minimal realization should be used for implementation.

Assumption 3.3. The operator has designed A_c, B_c, C_c, and D_c, such that the closed-loop system is asymptotically stable.

Assuming a stable closed-loop system is in line with normal operator requirements and is thus not restrictive.

Assumption 3.4. The closed-loop system dynamics have reached steady state at $k = 0$.

This assumption is not very restrictive, since industrial plants usually run for long periods of time. Due to Assumption 3.3 and the stationarity of $w(k)$ and $v(k)$, the closed-loop system will therefore converge to a stationary process.

3.2 Anomaly Detector Model

Under nominal conditions we have $\tilde{y}(k) = y(k)$ and $\tilde{u}(k) = u(k)$, which indicates that the plant is working as it is supposed to be. However, anomalies can occur at some a priori unknown point in time, like a sensor fault or a purposely injected signal by an attacker. Therefore, control systems need to detect these anomalies and make the operator aware of the fault by triggering an alarm. Since sensor attacks are of interest in this thesis, we make the following assumption.

Assumption 3.5. The only anomalies are the attack signals $y_a(k)$ in $\mathbb{R}^{n_y}$. This implies $\tilde{u}(k) = u(k)$ and $\tilde{y}(k) = y(k) + y_a(k)$, where $y_a(k)$ is designed by the attacker.

The anomaly detector computes a signal $y_D(k+1) \geq 0$ at time-step k, which is used to determine if an attacker is present or not. A small $y_D(k+1)$ indicates that no anomalies are present. If $y_D(k+1)$ grows large and crosses a threshold $J_D \geq 0$, an alarm is triggered. In case an alarm is triggered and no fault or intruder is present, we call it a *false alarm* and otherwise a *true alarm*. Typically, J_D is tuned such that rarely any false alarm happens. This means that a plant's operator will not be suspicious if there are no alarms happening for a longer period of time. The detector dynamics are described by a possibly nonlinear discrete-time system,

$$\begin{aligned} x_D(k+1) &= \theta\big(x_D(k), r(k)\big), \\ y_D(k+1) &= d\big(x_D(k), r(k)\big), \end{aligned} \tag{3.3}$$

where $x_D(k)$ is the internal state of the detector in $\mathbb{R}^{n_D}$, which is initialized as a zero vector, $y_D(k+1)$ is the output of the detector in $\mathbb{R}_{\geq 0}$, and $r(k)$ is the input to the detector in $\mathbb{R}^{n_r}$. Here, $\theta\big(x_D(k), r(k)\big)$ describes the dynamics of the detector state and $d\big(x_D(k), r(k)\big)$ is the output behavior of the detector. If the detector has no internal state, $x_D(k)$, we call it *stateless*, and *stateful* otherwise. The input of the detector is a random variable, whose distribution contains information about the status of the closed-loop system.

Assumption 3.6. The detector input $r(k)$ is a normalized residual signal that represents the difference between the received and the expected measurements, i.e. $r(k) = \Sigma_r^{-\frac{1}{2}}\big(\tilde{y}(k) - C\hat{x}(k)\big)$, where $C\hat{x}(k)$ is a prediction of the plant's output. Here, Σ_r is the covariance matrix of $\tilde{y}(k) - C\hat{x}(k)$ and the normalization with $\Sigma_r^{-\frac{1}{2}}$ leads to $r(k) \sim \mathcal{N}(0, I_{n_y})$ under nominal conditions.

Due to the normal distribution of the residual signal, $y_D(k+1)$ is also a random variable with probability density function $q_{k+1}(y_D)$ and support $\operatorname{supp}\big(q_{k+1}(y_D)\big) \subseteq [0, \infty)$, where $\operatorname{supp}\big(q(y_D)\big) := \{y_D \in \mathbb{R} : \ q(y_D) > 0\}$. It has been argued, for example in [19], that $x_D(k)$ is confidential and only the operator has access to it, since $x_D(k)$ is an internal value of the detector, which is not transmitted over a network. In Chapter 5, we show how an attacker can break the confidentiality of detector states under the assumption that follow linear dynamics. Additionally, the trajectory of $y_D(k)$ is displayed in the control center, such that an operator could recognize suspicious behavior by examining the displayed trajectory. This can also lead to the detection of the attacker and is represented as the human observer in Figure 3.1.

We constraint $\theta(x_D(k), r(k))$ and $d\big(x_D(k), r(k)\big)$ as follows to obtain a reasonable model of a detector in (3.3).

Assumption 3.7. The following conditions hold for the detector (3.3)

1) $\theta\big(x_D(k), r(k)\big)$ and $d\big(x_D(k), r(k)\big)$ are continuous in $x_D(k)$ and $r(k)$,

2) $y_D(k+1) = d\big(x_D(k), 0\big) \begin{cases} < J_D & \text{if } x_D(k) \neq 0 \text{ and } y_D(k) \leq J_D \\ = 0 & \text{if } x_D(k) = 0 \end{cases}$,

3) $\theta(0,0) = 0$ and $x_D(k) \to 0$ for $k \to \infty$, if $r(k) = 0 \quad \forall k$,

4) $d\big(x_D(k), r(k)\big)$ is coercive in $x_D(k)$ and $r(k)$,

5) Set $x_D(k) = 0$, if $y_D(k) > J_D$.

The first condition is needed for mathematical tractability. The second and third condition are needed to guarantee that if we have perfect predictions of the received measurements, i.e. $r(k) = 0$, the detector state and output will approach zero without causing a false alarm and $x_D(k) = 0$ is a global asymptotic equilibrium. For the fourth condition recall that a function $\alpha : \mathbb{R}^{n_D} \times \mathbb{R}^{n_y} \to \mathbb{R}_{\geq 0}$ is called coercive if $\alpha(x) \to \infty$ as $||x|| \to \infty$. Hence, the fourth condition guarantees that if

either the detector state or the detector input grow unbounded the output will also grow unbounded. The fifth condition means that the detector is reset to its initial state, when an alarm has been triggered. The reset is needed to avoid triggering false alarms consecutively.

3.2.1 Commonly-used Anomaly Detectors

Three commonly-used detectors that follow (3.3) and fulfill Assumption 3.7 are presented below. These detectors are the χ^2 detector, the CUSUM detector, and the MEWMA detector.

The χ^2 Detector

Since $r(k)$ is a residual signal that is determined by the difference between the received and expected measurement signal, it is reasonable to use the size of the residual signal as an indication of good plant behavior. The χ^2 detector is a stateless detector and looks at the size of the detector input by taking the squared Euclidean norm of $r(k)$,

$$y_D(k+1) = r(k)^T r(k). \tag{3.4}$$

If $y_D(k+1) > J_D^{\chi^2}$ an alarm is triggered, where $J_D^{\chi^2} \in \mathbb{R}_{\geq 0}$ is the chosen threshold of the χ^2 detector.

The MEWMA Detector

Instead of only looking at the size of $r(k)$, as in the χ^2 detector, the MEWMA detector [73] first filters $r(k)$ and then determines the size of the filtered signal. Due to the filter, the MEWMA detector is a stateful detector with the following dynamics,

$$x_D(k+1) = \beta r(k) + (1-\beta) x_D(k), \tag{3.5}$$

$$y_D(k+1) = \frac{2-\beta}{\beta} x_D(k+1)^T x_D(k+1), \tag{3.6}$$

where $x_D(k)$ is initialized as zero and $\beta \in (0, 1]$ is the forgetting factor of the detector. The threshold of the MEWMA detector is $J_D^M \in \mathbb{R}_{\geq 0}$ and an alarm is triggered if $y_D(k+1) > J_D^M$. Here, $\frac{2-\beta}{\beta}$ is a normalization factor for the MEWMA detector such that $y_D(k+1)$ under nominal conditions converges to a χ^2 distribution with n_y degrees of freedom. Note that for $\beta = 1$ the MEWMA detector represents the χ^2 detector.

The CUSUM Detector

The CUSUM detector does not filter the residual signals, but sums their squared Euclidean norm up with a forgetting factor to determine $y_D(k)$. The non-parametric

version of the CUSUM detector proposed in [74] is defined as follows

$$x_D(k+1) = \max\big(x_D(k) + r(k)^T r(k) - \delta, 0\big),$$
$$y_D(k+1) = x_D(k+1),$$

where $x_D(k)$ is initialized as zero and $\delta \in \mathbb{R}_{\geq 0}$ is the forgetting factor of the CUSUM detector. An alarm is triggered in case $y_D(k+1) > J_D^C$, where $J_D^C \in \mathbb{R}_{\geq 0}$ is the threshold of the CUSUM detector. Note that the internal state $x_D(k)$ is equal to the output of the detector in this case and, therefore, we write the detector dynamics as

$$y_D(k+1) = \max\big(y_D(k) + r(k)^T r(k) - \delta, 0\big).$$

Since $y_D(k)$ is used to calculate $y_D(k+1)$, this is also a stateful detector. The non-parametric CUSUM detector is also used in a similar fashion in, for example, [19] and [25].

3.2.2 Detector Comparison and Tuning

Before we move on to describe the attack model, we will discuss how we compare the anomaly detector in the presence of attacks and how to tune them, i.e. choose J_D.

Metrics to Compare Detectors

Above we presented a general detector model and three different anomaly detectors that fit into this detector model. A question that comes naturally to our mind is if it is possible to compare the performance of different detectors.

A common way to compare the performance of detectors is the receiver operating characteristic (ROC) curve [75]. The ROC curve plots the true alarm rate over the false alarm rate. The true alarm rate states how often the detector detects an attack, while the false alarm rate states how often the detector triggers an alarm without an attack being present. One way obtain the ROC curve is to tune J_D to achieve a certain false alarm rate and then the true alarm rate of the detector is determined for this threshold.

A sophisticated attacker typically designs the attack signal in such away that the detector will not trigger an alarm. In that case, the ROC curve cannot be used for performance comparison of detectors in the presence of attacks, because the detector will not trigger an alarm and the true alarm rate is always zero no matter which threshold is chosen. Instead of considering the true alarm rate, Urbina *et al.* [25] propose to examine the impact an attacker has on the closed-loop system while remaining undetected. The attack impact depends on the attacker's objective. The objective could, for example, be to drive the $x(k)$ as far away from the nominal state as possible. A more thorough definition of the attack impact can be found in Definition 3.2.

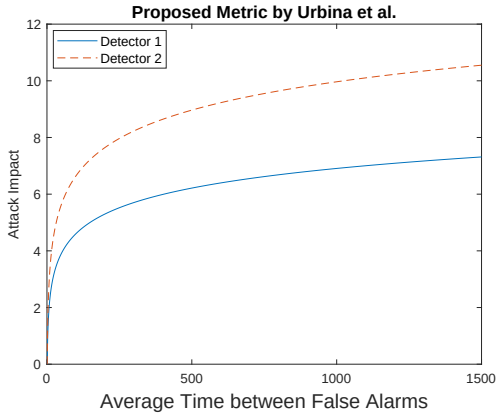

Figure 3.2: Comparing two detectors with the metric proposed in [25]

Furthermore, according to [25], it is also more intuitive for an operator to talk about the average time between false alarms instead of the false alarm rate. For example, telling someone that a false alarm on average happens every five weeks is easier to grasp then telling that 0.01% of alarms are false alarms. Interestingly, the mean time between false alarms is the inverse of the false alarm rate [76]. Figure 3.2 shows how this metric could look like when we compare two detectors. According to this metric, Detector 1 has a better performance than Detector 2 because using Detector 2 results in a higher attack impact than using Detector 1 for all investigated average times between false alarms.

In the case of randomly occurring faults, there is a trade-off between the true and false alarm rate we need to consider when choosing a detector. However, with attacks in mind, we should also consider the impact of stealthy attacks when choosing a detector.

Detector Tuning

Now that we know how to compare different detectors, we need to be able to obtain the metric. For that we need to tune the detector threshold J_D to achieve a certain mean time between false alarms.

Assumption 3.8. For a given detector (3.3), which fulfills the conditions in Assumption 3.7, there exists a compact set $\mathbb{L} \subseteq \mathbb{R}_{\geq 1}$ and a non-decreasing function $g : \mathbb{L} \to \mathbb{R}_{\geq 0}$ such that $J_D = g(\tau)$, where $\tau \in \mathbb{L}$ is the mean time between false alarms.

Assumption 3.8 is plausible, because an alarm is triggered when $y_D(k)$ crosses J_D. Therefore, a larger J_D means that a random input $y_D(k)$ needs more time until it crosses J_D. We need to introduce $\mathbb{L}$ because depending on the form of both $\theta\big(x_D(k), r(k)\big)$ and $d\big(x_D(k), r(k)\big)$ we might not be able to achieve any $\tau \in \mathbb{R}_{\geq 1}$ by

simply adjusting J_D. With Assumption 3.8 the detector tuning J_D can be specified by τ alone. How to choose J_D is a tedious task and depends on the what detector is used. Therefore, we will present how to tune J_D for the commonly used anomaly detectors.

Let us start with the χ^2 detector. The squared Euclidean norm of $r(k)$ has a χ^2 distribution with n_y degrees of freedom, since $r(k) \sim \mathcal{N}(0, I_{n_y})$. Hence, there exists a closed-form solution to obtain the threshold for a given τ,

$$J_D^{\chi^2} = 2P^{-1}\left(\tfrac{n_y}{2}, 1 - \tfrac{1}{\tau}\right), \tag{3.7}$$

where $P^{-1}(\cdot,\cdot)$ represents the inverse regularized lower incomplete gamma function (see Theorem 3 in [19]).

Now we look at the CUSUM and MEWMA detectors. Since both the MEWMA and the CUSUM detector have an internal state, the detector might be unstable. If the detectors are not stable, $y_D(k)$ might grow unbounded, even if the input $r(k)$ is bounded, and, thus, always hit the threshold J_D even if there is no fault. In [77] a notion of stochastic stability is introduced. However, we will only look at the boundedness, which is defined as follows.

Definition 3.1. A stochastic process $y_D(k)$ is bounded in mean square if

$$\mathbb{E}\{||y_D(k)||_2^2\} < \infty$$

for all $k \geq 0$.

Therefore, we start by investigating conditions for the MEWMA and CUSUM detector to be bounded in mean square.

Proposition 3.1. The CUSUM detector is bounded in mean square if $\delta > n_y$ and the MEWMA detector is bounded in mean square if and only if $\beta \in [0, 2)$.

Proof. Theorem 1 in [19] shows the boundedness condition for the CUSUM detector.

Since the MEWMA detector state is initialized with zero, i.e., $x_D(0) = 0$, the state at time step k is

$$x_D(k) = \beta \sum_{i=0}^{k-1} (1-\beta)^{k-1-i} r(i),$$

where we assumed that no reset has happened. We know that each $r(i)$ has a standard Gaussian distribution and is independent from all previous $r(j)$ with $j < i$. Therefore, we can determine the variance of $x_D(k)$ as

$$\begin{aligned}\operatorname{Var}\big(x_D(k)\big) &= \beta^2 \sum_{i=0}^{k-1} (1-\beta)^{2i} I_{n_y} = \beta^2 \frac{1-(1-\beta)^{2k}}{1-(1-\beta)^2} I_{n_y} \\ &= \frac{\beta}{2-\beta}\big(1-(1-\beta)^2\big) I_{n_y},\end{aligned}$$

such that $x_D(k) \sim \mathcal{N}\big(0, \beta\frac{1-(1-\beta)^{2k}}{2-\beta} I_{n_y}\big)$. From that we obtain

$$y_D(k) = \big(1-(1-\beta)^{2k}\big)\upsilon,$$

where υ has a χ^2 distribution with n_y degrees of freedom. Finally, using $\mathbb{E}\{y_D(k)^2\} = \mathbb{E}\{y_D(k)\}^2 + \mathrm{Var}\big(y_D(k)\big)$ leads to

$$\mathbb{E}\{y_D(k)^2 | y_D(0)=0\} = (2n_y + n_y^2)\big(1-(1-\beta)^{2k}\big)^2 < \infty \tag{3.8}$$

for all $k \geq 0$ if and only if $\beta \in [0,2)$. This is fulfilled for the MEWMA detector and leads to boundedness in mean square according to Definition 3.1. □

In contrast to the CUSUM detector, one does not have to worry about how to choose the forgetting factor for the MEWMA detector when it comes to stochastic boundedness. Although there is no upper bound for δ, we show in Chapter 6 that the attack impact grows with δ. Hence, the operator should not choose δ too large.

Now that we know what values for the forgetting factors need to be respected, let us look at determining the threshold for the stateful detectors. Determining J_D^C exactly is not an easy task, because there exists no closed-form solution as for $J_D^{\chi^2}$, but one can approximate J_D^C by approximating the continuous CUSUM scheme with an absorbing Markov chain with $R+1$ states [19]. With this method, one can approximate τ for a given threshold J_D^C, but also find a threshold when τ is given using a bisection method. Note that, the computed threshold J_D^C will only approximately achieve the desired τ, but as $R \to \infty$ the approximation approaches the real solution (see Theorem 2 and Remark 2 of [19] for more details).

Similarly to [19], the MEWMA detector is approximated with an absorbing Markov chain with $R+1$ states in [78] to approximate τ for a given β and J_D^M. Therefore, we are also able to approximate J_D^M for a given τ and β using the Markov chain and a bisection method.

3.3 Attack Model

After introducing both the plant and controller model as well as a general detector model with tuning methods and comparison metrics, we present now the model of the attacker.

Assumption 3.9. The attacker has gained access to the plant model (A, B, C), the controller model (A_c, B_c, C_c, D_c), the detector used, its threshold J_D, the noise statistics (Σ_w, Σ_v), the measurements $y(k)$ for $k \geq 0$ but *not* the control signals $u(k)$, the initial state of the plant $x(0)$, controller $x_c(0)$, and detector $x_D(0)$.

Since the manipulation of control signals can lead to an immediate physical impact, we assume $u(k)$ is better protected and, therefore, the attacker does not have access to it. Moreover, we set the start of the attack arbitrarily to $k = 0$. This is interpreted as the point in time, from which the attacker has access to the

measurements. From Assumption 3.4 we know that the plant and controller have been running for a long time. Therefore, the attacker cannot know $x(0)$, $x_c(0)$, and $x_D(0)$, when it gains access to the sensor measurements. The reason for that is that the states at $k = 0$ depend on the system's past inputs, initial states, and noise signals, which the attacker does not know.

Recall that $\tilde{y}(k) = y(k) + y_a(k)$ such that the closed-loop dynamics are

$$\begin{bmatrix} x(k+1) \\ x_c(k+1) \end{bmatrix} = \begin{bmatrix} A + BD_cC & BC_c \\ B_cC & A_c \end{bmatrix} \begin{bmatrix} x(k) \\ x_c(k) \end{bmatrix} + \begin{bmatrix} BD_c \\ B_c \end{bmatrix} y_a(k) + \begin{bmatrix} w(k) + BD_cv(k) \\ B_cv(k) \end{bmatrix}.$$

By introducing

$$z(k) = \begin{bmatrix} x(k) \\ x_c(k) \end{bmatrix} \text{ and } \eta'(k) = \begin{bmatrix} w(k) + BD_cv(k) \\ B_cv(k) \end{bmatrix}$$

we rewrite the closed-loop system as

$$\begin{aligned} z(k+1) &= A'_z z(k) + B_z y_a(k) + \eta'(k), \\ y(k) &= C_z z(k) + y_a(k) + v(k) = \begin{bmatrix} C & 0 \end{bmatrix} z(k) + y_a(k) + v(k), \end{aligned} \tag{3.9}$$

where $\eta'(k) \sim \mathcal{N}(0, Q')$ is the zero mean process noise of the closed-loop system with covariance matrix $Q' \in \mathbb{R}^{(n_x+n_c)\times(n_x+n_c)}$ and $v(k)$ is the measurement noise. Note that although we only investigate sensor attacks, the attack can have a direct influence on the actuator if the controller contains a non-zero feedthrough term. Due to Assumption 3.3 and 3.4 and the Gaussian noise processes, we know that $\rho(A'_z) < 1$ and that $z(0) \sim \mathcal{N}(0, \Sigma_0)$, where Σ_0 is the unique solution to

$$\Sigma_0 = A'_z \Sigma_0 (A'_z)^T + Q'.$$

Since the closed-loop system is linear, we can split it into two parts, such that $z(k) = z_n(k) + z_a(k)$, where

$$z_n(k+1) = A'_z z_n(k) + \eta'(k) \text{ and } z_a(k+1) = A'_z z_a(k) + B_z y_a(k) \tag{3.10}$$

with $z_n(0) = z(0)$ and $z_a(0) = 0$. Here, $z_n(k)$ represents the part of $z(k)$ that is excited by the noise, while $z_a(k)$ is the part that is excited by the attack signal. Further, since $\mathbb{E}\{z(0)\} = 0$ and $\mathbb{E}\{\eta'(k)\} = 0$ for all k, we can interpret $z_a(k)$ as the mean of the closed-loop system.

3.3.1 Worst-case Attack Strategy and Its Impact

Now that we determined the closed-loop system, we present a worst-case attack strategy and the definition of the worst-case attack impact. However, we show that to execute the worst-case attack strategy, the attacker needs to gather more knowledge than what is assumed in Assumption 3.9.

First, we make the following assumption on the design of $y_a(k)$.

Assumption 3.10. The attacker designs $y_a(k)$ such that the attack remains undetected, i.e.

$$y_D(k) = d\big(x_D(k), r(k)\big) \leq J_D \quad \forall k \geq 0.$$

To fulfill Assumption 3.10 let us look at the input of the detector and the worst-case attack strategy. The input to the detector is

$$r(k) = \Sigma_r^{-\frac{1}{2}}\big(y(k) - C\hat{x}(k) + y_a(k)\big),$$

so Murguia *et al.* [19] propose the following attack strategy

$$y_a(k) = -y(k) + C\hat{x}(k) + \Sigma_r^{\frac{1}{2}} a(k), \tag{3.11}$$

where $a(k) \in \mathbb{R}^{n_y}$ is a vector chosen by the attacker. This attack strategy is a worst-case attack strategy, because the detector input becomes $r(k) = a(k)$. Therefore, the attacker has full control over the input of the detector, which makes it easier for the attacker to remain undetected, i.e. design $a(k)$ such that

$$y_D(k+1) = d\big(x_D(k), a(k)\big) \leq J_D \quad \forall\, k \geq 0.$$

The closed-loop system dynamics under this attack become

$$\begin{bmatrix} x(k+1) \\ x_c(k+1) \end{bmatrix} = \begin{bmatrix} A & B\big(C_c + D_c C T_c\big) \\ 0 & A_c + B_c C T_c \end{bmatrix} \begin{bmatrix} x(k) \\ x_c(k) \end{bmatrix} + \begin{bmatrix} BD_c \\ B_c \end{bmatrix} \Sigma_r^{\frac{1}{2}} a(k) + \begin{bmatrix} w(k) \\ 0 \end{bmatrix}, \tag{3.12}$$

where we used that $\hat{x}(k) = T_c x_c(k)$ (Assumption 3.1).

However, an attacker with the model knowledge according to Assumption 3.9 will not be able to execute the attack strategy in (3.11). The reason for this is that the attacker has no access to $x_c(k)$ when the attack starts. Therefore, it cannot determine $\hat{x}(k)$ to execute the attack strategy in (3.11). Further, even if the attacker has access to $\hat{x}(k)$ without access to $x_D(k)$, the attacker needs to apply a conservative attack strategy to remain undetected (see Chapter 6). Therefore, utilizing stateful detectors increase the security of the system, if the attacker is not able to obtain $x_D(k)$ during the execution of its attack.

Therefore, it is of interest to investigate if an attacker with the knowledge according to Assumption 3.9 is able to launch the stealthy worst-case attack (3.11). To launch this attack the attacker has to first find a way to obtain $x_c(k)$ and second to acquire $x_D(k)$ if the attacker wants to maximize its impact and a stateful detector is used. Chapter 4 investigates when an attacker is able to get a perfect estimate of $x_c(k)$, while Chapter 5 examines how an attacker can estimate $x_D(k)$.

Once the attacker can launch the attack proposed in (3.11), we can determine the worst-case impact of this attack when a certain detector is used. Further, for mathematical tractability, we also assume the following.

Assumption 3.11. The attack (3.11) is time-limited to the interval $\Gamma \in [\underline{k}, \bar{k}]$, i.e. the attack starts at time step $\underline{k}$ and ends at $\bar{k}$.

This means that from $k = 0$ to $k = \underline{k} - 1$ the attacker obtains both $x_c(k)$ and $x_D(k)$ and then launches the attack (3.11) at $k = \underline{k}$. The conditions for the attack to remain undetected are then given by

$$d\big(x_D(k), a(k)\big) \leq J_D \quad \forall\, k \in \Gamma. \tag{3.13}$$

Assumption 3.11 leads to the definition of a worst-case impact of a time-limited attack.

Definition 3.2. The worst-case impact $\mathcal{I} : \mathbb{L} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ of the stealthy sensor attack (3.11) on the closed-loop system (3.12) with zero initial conditions, no noise ($w(k) = 0$), and equipped with an anomaly detector (3.3) is defined as

$$\mathcal{I}(\tau) := \max_{a, x_D(\underline{k})} f(a) \quad \text{s.t.} \quad d\big(x_D(k), a(k)\big) \leq J_D \quad \forall\, k \in \Gamma,$$

where $a = \{a(k)\}_{k=\underline{k}}^{\bar{k}}$ is the attack trajectory and $f(a)$ is a continuous function that characterizes the attacker's objective.

Recall that $J_D = g(\tau)$ (Assumption 3.8) and, therefore, the impact depends on τ. here, the attacker's objective $f(a)$ could, for example, be monetary loss or physical damage to the system. From the perspective of a defender, the defender does not know when the attack will happen and, therefore, does not know $x_D(\underline{k})$, which is the state of the detector at the beginning of the attack given by (3.11). Hence, we optimize over all possible $x_D(\underline{k})$ to find the worst-case impact. Further, $\mathcal{I}(\tau) = \infty$ means that the impact of the stealthy attack is unbounded. Note that plotting $\mathcal{I}(\tau)$ over τ for different detectors, gives us the metric for detector comparison we mentioned earlier.

Throughout the course of this thesis, we will often set the attacker's objective to $f(a) = ||T_a a||_\infty$. Let us discuss the two reasons why this attack objective is used. Due to the linearity of $x_a(k)$, we can express many different quantities in the system by $T_a a$ such as critical states at a specific time step or whole trajectories. One example is $x_a(\bar{k}) = T_a a$, where $T_a \in \mathbb{R}^{n_x \times (\bar{k} - \underline{k} + 1) n_y}$. In that way, maximizing $||T_a a||_\infty$ maximizes the largest element in $x_a(\bar{k})$. This can for example be interpreted as pressure in a tank that the attacker wants to maximize, in order to make the tank explode. The second reason for choosing this attack objective is that

$$\max_a ||T_a a||_\infty = \max_a \max_{i \in \{1, \cdots, n_a\}} |t_{a,i}^T a|,$$

where $t_{a,i} \in \mathbb{R}^{(\bar{k} - \underline{k} + 1) n_y}$ is the ith row of $T_a \in \mathbb{R}^{n_a \times (\bar{k} - \underline{k} + 1) n_y}$. By splitting the maximization of $||T_a a||_\infty$ into n_a subproblems we are able to find analytical solutions and global optima for nonconvex problems (see Chapter 5 and 6).

3.4 Observer-based Controllers under Attack

Since all of the illustrative examples in this thesis use an observer-based controller [79], we present the closed-loop dynamics of an observer-based controller under the sensor attack (3.11) in this section.

An observer-based controller uses an estimate $\hat{x}(k)$ of the plant's state $x(k)$ to determine the control input. In this case, we have $x_c(k) = \hat{x}(k)$ such that Assumption 3.1 is fulfilled. The dynamics of this controller are

$$\begin{aligned}\hat{x}(k+1) &= (A - BK - LC)\hat{x}(k) + LCy(k),\\ u(k) &= -K\hat{x}(k),\end{aligned}$$

where K is the controller gain and L is the observer gain. The closed-loop system is then given by

$$\begin{bmatrix} x(k+1) \\ \hat{x}(k+1) \end{bmatrix} = \begin{bmatrix} A & -BK \\ LC & A - BK - LC \end{bmatrix} \begin{bmatrix} x(k) \\ \hat{x}(k) \end{bmatrix} + \begin{bmatrix} 0 \\ L \end{bmatrix} y_a(k) + \begin{bmatrix} w(k) \\ Lv(k) \end{bmatrix}.$$

With the introduction of the estimation error $e(k) = x(k) - \hat{x}(k)$, we can write the closed-loop system as

$$\begin{bmatrix} x(k+1) \\ e(k+1) \end{bmatrix} = \begin{bmatrix} A - BK & BK \\ 0 & A - LC \end{bmatrix} \begin{bmatrix} x(k) \\ e(k) \end{bmatrix} + \begin{bmatrix} 0 \\ -L \end{bmatrix} y_a(k) + \begin{bmatrix} w(k) \\ w(k) - Lv(k) \end{bmatrix}. \tag{3.14}$$

From (3.14), we see that K and L need to be designed such that $\rho(A - BK) < 1$ and $\rho(A - LC) < 1$ to fulfil Assumption 3.3.

The normalized residual signal with $e(k)$ is given by

$$\begin{aligned} r(k) &= \Sigma_r^{-\frac{1}{2}} \big(\tilde{y}(k) - C\hat{x}(k)\big) = \Sigma_r^{-\frac{1}{2}} \big(y(k) + y_a(k) - C\hat{x}(k)\big)\\ &= \Sigma_r^{-\frac{1}{2}} \big(Ce(k) + v(k) + y_a(k)\big), \end{aligned}$$

such that the worst-case attack (3.11) becomes

$$y_a(k) = -y(k) + C\hat{x}(k) + \Sigma_r^{\frac{1}{2}} a(k) = -Ce(k) - v(k) + \Sigma_r^{\frac{1}{2}} a(k).$$

The closed-loop dynamics with the worst-case attack are

$$\begin{bmatrix} x(k+1) \\ e(k+1) \end{bmatrix} = \begin{bmatrix} A - BK & BK \\ 0 & A \end{bmatrix} \begin{bmatrix} x(k) \\ e(k) \end{bmatrix} + \begin{bmatrix} 0 \\ -L\Sigma_r^{\frac{1}{2}} \end{bmatrix} a(k) + \begin{bmatrix} w(k) \\ w(k) \end{bmatrix}. \tag{3.15}$$

Here, we see that if $\rho(A) > 1$ the error dynamics diverge even if $a(k) = 0$ for all k. This means if the plant itself has unstable dynamics the worst-case attack will make the closed-loop system unstable.

3.5 Summary

In this chapter, we presented our model for the sensor attack. We started by introducing the models of the linear plant and controller and then presented a general detector model. Three detectors, namely the χ^2, CUSUM, and MEWMA detectors, that fit into the detector model were presented. Then we introduced an attacker model, which has full model knowledge and access to all sensors, and a worst-case attack strategy. For its execution, this attack strategy needs access to the controller state. We argued that it is not possible for the attacker to have access to the controller state, when the attack starts. Therefore, the next step is to investigate if an attacker according to our model is able to access the controller state by eavesdropping on the measurements.

Chapter 4

On the Confidentiality of the Controller State

4.1 Problem Formulation

In Chapter 3, we introduced the worst-case attack strategy (3.11) and discussed that the attacker is not able to launch this attack strategy until it has obtained a perfect estimate of $\hat{x}(k)$. Since $\hat{x}(k)$ is included in the controller's state $x_c(k)$ (Assumption 3.1), this chapter investigates if an attacker according to Assumption 3.9 is able to estimate $x_c(k)$ by eavesdropping on the measurements $y(k)$. It might seem obvious that an attacker that knows the model and the measurements can estimate $x_c(k)$. This is, however, not always true and we present necessary and sufficient conditions for when the attacker is able to perfectly estimate the controller's state. Furthermore, we present a defense mechanism and verify the results with a simulation of a three tank system.

We consider the confidentiality attack in this chapter as a first step for the attacker to be able to execute (3.11). However, we can also interpret the attack as an attack on the privacy of the controller state. In that case, the attacker is not malicious but curious. In this chapter, the attacker does not use its disruptive resources, i.e. $y_a(k) = 0$. If an alarm is triggered, it is thus triggered by the noise and not by an attack signal. However, since the attacker is in the network, this alarm is not a false alarm. Therefore, the alarms could lead to a detection of the attacker. For now, though, we do not consider the detection of the attacker by these alarms and only look at the feasibility of estimating $x_c(k)$ perfectly.

Let us first restate the closed-loop dynamics to set the stage to formulate the problem investigated in this chapter.

Recall that the closed-loop system with $y_a(k) = 0$ is given by

$$\begin{aligned} z(k+1) &= A_z' z(k) + \eta'(k) \\ y(k) &= C_z z(k) + v(k), \end{aligned} \tag{4.1}$$

where $z(k) = [x(k)^T \; x_c(k)^T]$, $\eta'(k) \sim \mathcal{N}(0, Q')$ is the zero mean process noise of the

closed-loop system with covariance matrix $Q' \in \mathbb{R}^{(n_x+n_c)\times(n_x+n_c)}$, and $v(k)$ is the measurement noise.

Note that the closed-loop process noise variable $\eta'(k)$ is correlated with the measurement noise $v(k)$,

$$\mathbb{E}\left\{ \begin{bmatrix} \eta'(k) \\ v(k) \end{bmatrix} \begin{bmatrix} \eta'(k)^T & v(k)^T \end{bmatrix} \right\} = \left[\begin{array}{cc|c} \Sigma_w + BD_c\Sigma_v D_c^T B^T & BD_c\Sigma_v B_c^T & BD_c\Sigma_v \\ B_c\Sigma_v D_c^T B^T & B_c\Sigma_v B_c^T & B_c\Sigma_v \\ \hline \Sigma_v B^T D_c^T & \Sigma_v^T B_c^T & \Sigma_v \end{array}\right] = \left[\begin{array}{c|c} Q' & S \\ \hline S^T & R \end{array}\right],$$

where $S \in \mathbb{R}^{(n_x+n_c)\times n_y}$, and $R \in \mathbb{R}^{n_y \times n_y}$.

Since the $\eta'(k)$ and $v(k)$ are correlated, we will apply a transformation proposed in [80] to obtain a system representation with uncorrelated noises,

$$\begin{aligned} z(k+1) &= A_z' z(k) + \eta'(k) - SR^{-1}(y(k) - y(k)) \\ &= A_z z(k) + \eta(k) + SR^{-1} y(k), \end{aligned}$$

where $A_z = A_z' - SR^{-1}C_z$,

$$\eta(k) = \eta'(k) - SR^{-1}v(k) = \begin{bmatrix} w(k) \\ 0 \end{bmatrix},$$

$$\mathbb{E}\left\{ \begin{bmatrix} \eta(k) \\ v(k) \end{bmatrix} \begin{bmatrix} \eta(k)^T & v(k)^T \end{bmatrix} \right\} = \left[\begin{array}{c|c} Q & 0 \\ \hline 0 & R \end{array}\right],$$

and

$$Q = Q' - SR^{-1}S^T = \begin{bmatrix} \Sigma_w & 0 \\ 0 & 0 \end{bmatrix}.$$

The zero elements in Q show us that there is no process noise acting on the controller state in the transformed system.

Therefore, the closed-loop dynamics we consider in this chapter are

$$\begin{aligned} z(k+1) &= A_z z(k) + \eta(k) + SR^{-1}y(k), \\ y(k) &= C_z z(k) + v(k). \end{aligned} \tag{4.2}$$

Note that even though $\rho(A_z') < 1$, it is not always the case that $\rho(A_z) < 1$.

Before we present the problem formulation of this chapter, we make the following assumption on the attacker.

Assumption 4.1. The attacker uses measurements up to time step k to estimate the controller's internal state at time step $k+1$.

It is possible to use measurements up to time step $k^* \geq k+1$ to estimate the controller's state at time step $k+1$. However, if the attacker wants to launch the worst-case attack (3.11) at time step $k+1$, this estimate needs to be available already.

In this chapter, the *goal* of the attacker is to obtain an estimate $\hat{x}_c(k)$, such that this estimate perfectly tracks the controller state $x_c(k)$ as k grows large. The goal can be formulated as the following problem.

Problem 4.1. Estimate $x_c(k)$ such that the estimation error is unbiased, i.e. $\mathbb{E}\{x_c(k) - \hat{x}_c(k)\} = 0$, and its covariance matrix $\Sigma_c(k)$ approaches zero, i.e.

$$\lim_{k\to\infty} \Sigma_c(k) = 0$$

for a given $\Sigma_c(0) \geq 0$.

An estimation error covariance matrix $\Sigma_c(k)$ that approaches zero as k grows large means the estimate converges to the true value in mean square (and thus also in probability).

In Section 4.2 we characterize for which systems the controller's confidentially can be broken (Problem 4.1), and in Section 4.3 we propose a defense mechanism and discus unstable controllers as a defense mechanism.

4.2 Estimating the Controller's State $x_c(k)$

In this section, we investigate when a solution to Problem 4.1 exists. It may seem obvious that an attacker according to Assumption 3.9 is without any doubt able to estimate the controller's state $x_c(k)$ perfectly. However, we show in the following that this is not always the case. First, we present the optimal attack strategy to estimate $x_c(k)$ and then state conditions for the convergence of $\Sigma_c(k)$ to zero. Following this, we look into non-optimal strategies to solve Problem 4.1.

4.2.1 Optimal Attack Strategy

To obtain the optimal attack strategy, we start by investigating the conditional probability of the closed-loop system state $z(k+1)$ given all measurements up to time step k. Due to the presence of the process noise, $\eta(k)$, and measurement noise, $v(k)$, we know that $z(k+1)$ is a random variable. Since (4.2) is a linear system with Gaussian noise, we know that $z(k+1)$ given the measurements up to time step k is also a Gaussian random variable [81]. Let $\{y(i)\}_{i=0}^{l}$ be the sequence $\{y(0), \cdots, y(l)\}$, then the conditional probability distribution of $z(k+1)$ given $\{y(i)\}_{i=0}^{k}$ is

$$z(k+1|\{y(i)\}_{i=0}^{k}) \sim \mathcal{N}\big(\hat{z}(k+1), \Sigma_z(k+1)\big),$$

where

$$\hat{z}(k+1) = A_z\hat{z}(k) + SR^{-1}y(k) + L_z(k)\big(y(k) - C_z\hat{z}(k)\big) \tag{4.3}$$

is the conditional mean of $z(k+1)$ with $L_z(k) = \big(A_z\Sigma_z(k)C_z^T\big)\big(C_z\Sigma_z(k)C_z^T + R\big)^{-1}$, $\hat{z}(0) = \mathbb{E}\{z(0)\} = 0$, and

$$\begin{aligned}\Sigma_z(k+1) &= A_z\Sigma_z(k)A_z^T + Q \\ &- \big(A_z\Sigma_z(k)C_z^T\big)\big(C_z\Sigma_z(k)C_z^T + R\big)^{-1}\big(A_z\Sigma_z(k)C_z^T\big)^T\end{aligned} \tag{4.4}$$

is the conditional covariance matrix. Its initial condition is $\Sigma_z(0) = \Sigma_0$, which is given in Assumption 3.4.

The optimal estimator for $z(k)$ given $\{y(i)\}_{i=0}^k$ is the Kalman filter [81]. It is optimal in the sense that it minimizes the mean square error. Therefore, the *optimal* attack strategy to estimate $x_c(k)$ is a time-varying Kalman filter, which uses $\hat{z}(k)$ in (4.3) as the estimate of $z(k)$. The goal of the attacker is to have an estimate $\hat{z}(k)$ of the closed-loop system's state such that $\begin{bmatrix}0 & I_{n_x}\end{bmatrix}\hat{z}(k) \to x_c(k)$ as $k \to \infty$.

Instead of directly analyzing $\hat{z}(k)$, we introduce the estimation error $e_z(k) = z(k) - \hat{z}(k)$ that has the dynamics

$$e_z(k+1) = \big(A_z - L_z(k)C_z\big)e_z(k) + \eta(k) + L_z(k)v(k).$$

and covariance matrix

$$\mathbb{E}\{e_z(k+1)e_z(k+1)^T|\{y(i)\}_{i=0}^k\big\} = \Sigma_z(k+1).$$

A Kalman filter is an unbiased estimator, which means that $\mathbb{E}\{z(k)\} = \hat{z}(k)$, or, differently formulated, $\mathbb{E}\{e_z(k)\} = 0$. Hence, Problem 4.1 is solved if, for $\Sigma_z(0) = \Sigma_0$, the attacker's Kalman filter fulfills

$$\lim_{k\to\infty}\Sigma_z(k) = \begin{bmatrix}P & 0\\ 0 & 0\end{bmatrix}, \tag{4.5}$$

where $P \geq 0$. Note that Σ_0 can be calculated by the attacker because of its model knowledge by Assumption 3.9.

4.2.2 Asymptotic Convergence to $\Sigma_c(k) = 0$

Let us now investigate when the optimal attack strategy solves Problem 4.1. Here, we present necessary and sufficient conditions for the covariance matrix $\Sigma_c(k)$ to converge to zero. Recall this is equivalent to saying that (4.5) is fulfilled.

Before we present our convergence results, note that a steady state solution to (4.4) satisfies the algebraic Riccati equation (ARE)

$$\Sigma_\infty = A_z\Sigma_\infty A_z^T + Q - \big(A_z\Sigma_\infty C_z^T\big)\big(C_z\Sigma_\infty C_z^T + R\big)^{-1}\big(A_z\Sigma_\infty C_z^T\big)^T, \tag{4.6}$$

where $L_\infty = \big(A_z\Sigma_\infty C_z^T\big)\big(C_z\Sigma_\infty C_z^T + R\big)^{-1}$ is the steady state Kalman gain.

Definition 4.1 (Definition 3.1 [80])**.** A real symmetric nonnegative definite solution Σ_∞ to (4.6) is called a *strong* solution if $\rho(A_z - L_\infty C_z) \leq 1$. The strong solution is called a *stabilizing* solution if $\rho(A_z - L_\infty C_z) < 1$.

The following lemma from [82] will be useful in the following discussion.

Lemma 4.1 (Theorem 3.2 [82])**.** *Let $G^T G = Q$,*

1. *the strong solution of the ARE exists and is unique if and only if (C_z, A_z) is detectable;*
2. *the strong solution is the only nonnegative definite solution of the ARE if and only (C_z, A_z) is detectable and (A_z, G) has no uncontrollable modes outside the unit circle;*
3. *the strong solution coincides with the stabilizing solution if and only if (C_z, A_z) is detectable and (A_z, G) has no uncontrollable modes on the unit circle;*
4. *the stabilizing solution is positive definite if and only if (C_z, A_z) is detectable and (A_z, G) has no uncontrollable modes inside, or on the unit circle.*

Let us begin by showing that a solution to (4.6) of the form in (4.5) exists.

Proposition 4.1. A solution of the algebraic Riccati equation (4.6) is given by

$$\Sigma_\infty = \begin{bmatrix} P & 0 \\ 0 & 0 \end{bmatrix},$$

where $P \geq 0$ is the unique strong solution of the ARE

$$P = APA^T + \Sigma_w - APC^T(CPC^T + \Sigma_v)^{-1}CPA^T.$$

Proof. Let us first determine

$$A_z = A_z' - SR^{-1}C_z = \begin{bmatrix} A & C_c \\ 0 & A_c \end{bmatrix}.$$

After algebraic computations we obtain

$$A_z\Sigma_\infty A_z^T + Q = \begin{bmatrix} APA^T + \Sigma_w & 0 \\ 0 & 0 \end{bmatrix}, A_z\Sigma_\infty C_z^T = \begin{bmatrix} APC^T \\ 0 \end{bmatrix},$$

and $C_z\Sigma_\infty C_z^T + R = CPC^T + \Sigma_v$ such that

$$\left(A_z\Sigma_\infty C_z^T\right)\left(C_z\Sigma_\infty C_z^T + R\right)^{-1}\left(A_z\Sigma_\infty C_z^T\right)^T = \begin{bmatrix} APC^T(CPC^T + \Sigma_v)^{-1}CPA^T & 0 \\ 0 & 0 \end{bmatrix}.$$

This leads to

$$\Sigma_\infty = \begin{bmatrix} APA^T + \Sigma_w - APC^T(CPC^T + \Sigma_v)^{-1}CPA^T & 0 \\ 0 & 0 \end{bmatrix}.$$

For Σ_∞ to be a solution of (4.6) we require

$$P = APA^T + \Sigma_w - APC^T(CPC^T + \Sigma_v)^{-1}CPA^T. \tag{4.7}$$

Note that (4.7) by itself is an algebraic Riccati equation. It is actually the algebraic Riccati equation an operator would obtain when it is designing a time-invariant Kalman filter for the plant's state. Due to the detectability of (C, A) (Assumption 3.2), there exists a unique strong solution $P \geq 0$ for (4.7) (Lemma 4.1). Hence, Σ_∞ is a solution of (4.6). □

Now that we proved that Σ_∞ is indeed a solution to the algebraic Riccati equation, we need to show under which conditions $\Sigma_z(k)$ converges to Σ_∞ for the initial condition Σ_0.

Lemma 4.2. *The unique strong solution of the ARE* (4.6) *is* Σ_∞ *if and only if* $\rho(A_c) \leq 1$.

Proof. Due to the first statement in Lemma 4.1, the strong solution is unique and exists if and only if (C_z, A_z) is detectable. From the stability of $A_z' = A_z + SR^{-1}C_z$ (see Assumption 3.3), it follows that (C_z, A_z) is detectable. Hence, the strong solution will be unique. Further, if $\rho(A_z - L_\infty C_z) \leq 1$ for

$$L_\infty = \left(A_z\Sigma_\infty C_z^T\right)\left(C_z\Sigma_\infty C_z^T + R\right)^{-1} = \begin{bmatrix} APC^T(CPC^T + \Sigma_v)^{-1} \\ 0 \end{bmatrix} = \begin{bmatrix} \bar{L} \\ 0 \end{bmatrix},$$

then Σ_∞ is a strong solution. Let us now look at the eigenvalues of $A_z - L_\infty C_z$, which are determined by the eigenvalues of $A - \bar{L}C$ and A_c, because

$$A_z - L_\infty C_z = \begin{bmatrix} A - \bar{L}C & C_c \\ 0 & A_c \end{bmatrix}.$$

Due to the detectability of (C, A) (Assumption 3.2), the first statement of Lemma 4.1 shows us that P is a strong solution of (4.7), such that $\rho(A - \bar{L}C) \leq 1$. Therefore, $\rho(A_z - L_\infty C_z) \leq 1$, i.e. Σ_∞ is the unique strong solution, if and only if $\rho(A_c) \leq 1$. □

Theorem 4.1. *The covariance matrix* $\Sigma_z(k)$ *converges to the attacker's desired covariance matrix* Σ_∞ *for the initial condition* Σ_0*, if and only if* $\rho(A_c) \leq 1$.

Proof. By Lemma 4.2, Σ_∞ is the unique strong solution of (4.6) if and only if $\rho(A_c) \leq 1$. Theorem 4.2 in [82] states that subject to $\Sigma_0 - \Sigma_\infty \geq 0$ the covariance matrix $\Sigma_z(k)$ will converge to the strong solution Σ_∞ if and only if (C_z, A_z) is

detectable. That (C_z, A_z) is detectable is shown in the proof of Lemma 4.2. Let us now show that $\Sigma_0 - \Sigma_\infty \geq 0$. If we use the system representation with correlated noise processes (4.1), the ARE for Σ_∞, according to [81], is

$$\Sigma_\infty = A_z' \Sigma_\infty (A_z')^T + Q' - (A_z' \Sigma_\infty C_z^T + S)(C_z \Sigma_\infty C_z^T + R)^{-1}(A_z' \Sigma_\infty C_z^T + S)^T. \tag{4.8}$$

Subtracting (4.8) from the Lyapunov equation for Σ_0 in Assumption 3.4 leads to

$$\Sigma_0 - \Sigma_\infty = A_z' \big(\Sigma_0 - \Sigma_\infty\big)(A_z')^T + (A_z' \Sigma_\infty C_z^T + S)(C_z \Sigma_\infty C_z^T + R)^{-1}(A_z' \Sigma_\infty C_z^T + S)^T.$$

This is also a Lyapunov equation with a unique solution since $\rho(A_z') < 1$ (Assumption 3.3). Further, we observe that

$$(A_z' \Sigma_\infty C_z^T + S)(C_z \Sigma_\infty C_z^T + R)^{-1}(A_z' \Sigma_\infty C_z^T + S)^T \geq 0,$$

because $\Sigma_\infty \geq 0$. Therefore, we know that $\Sigma_0 - \Sigma_\infty \geq 0$. Hence, with initial condition Σ_0

$$\lim_{k \to \infty} \Sigma_z(k) = \Sigma_\infty$$

if and only if $\rho(A_c) \leq 1$. □

Corollary 4.1. *Problem 4.1 is solvable if and only if* $\rho(A_c) \leq 1$.

Note that since the attacker uses a Kalman filter, it does not only obtain a perfect estimate of $x_c(k)$ but also an optimal estimate of $x(k)$.

Theorem 4.1 shows that the covariance matrix converges to the attacker's desired strong solution, but not how fast the convergence is. Therefore, we will now investigate the conditions for an exponential convergence rate.

Proposition 4.2. Subject to $\Sigma_0 > 0$, the covariance matrix $\Sigma_z(k)$ converges exponentially fast to Σ_∞ if and only if $\rho(A_c) < 1$.

Proof. Theorem 4.1 in [82] shows us that subject to $\Sigma_0 > 0$ the covariance matrix $\Sigma_z(k)$ converges exponentially fast to the stabilizing solution if and only if (C_z, A_z) is detectable and (A_z, G) has no uncontrollable modes on the unit circle. We already showed that (C_z, A_z) is detectable, therefore we look at the controllable modes of (A_z, G) now. Recall that $GG^T = Q$ such that

$$G = \begin{bmatrix} \Sigma_w^{\frac{1}{2}} & 0 \\ 0 & 0 \end{bmatrix}.$$

For (A_z, G) to have no uncontrollable modes on the unit circle we need A_c to have no eigenvalues on the unit circle, because we cannot control the eigenvalues of A_c with G, and due to Assumption 3.2 $(A, \Sigma_w^{\frac{1}{2}})$ has no uncontrollable modes on the unit circle. We showed in Lemma 4.2 that Σ_∞ is a strong solution to the ARE if and only if $\rho(A_c) \leq 1$. Hence, subject to $\Sigma_0 > 0$ the covariance matrix $\Sigma_z(k)$ converges exponentially fast to Σ_∞ if and only if $\rho(A_c) < 1$. □

This shows us that if $\Sigma_0 > 0$ and the operator uses a stable controller, i.e. $\rho(A_c) < 1$, the covariance matrix of the attacker's time-varying Kalman filter will converge exponentially fast to Σ_∞. Hence, the attacker is able to obtain a perfect estimate of $x_c(k)$ exponentially fast.

4.2.3 Breaking Confidentiality of $x_c(k)$ Using Non-optimal Observers

Previously, we have shown under which conditions the attacker is able to get a perfect estimate of the controller state $x_c(k)$ when a time-varying Kalman filter is used. The time-varying Kalman filter is the optimal filter for linear systems with Gaussian noise. One may wonder whether or not the attacker is able to perfectly estimate $x_c(k)$, when the attacker uses a non-optimal observer. Here, we investigate a time-invariant observer of the form

$$\hat{z}(k+1) = A_z\hat{z}(k) + SR^{-1}y(k) + L_z\big(y(k) - C_z\hat{z}(k)\big), \tag{4.9}$$

with $\hat{z}(0) = 0$, where L_z is the attacker's constant observer gain. As before, instead of looking at $\hat{z}(k)$, we analyze the error dynamics given by

$$e_z(k+1) = \big(A_z - L_zC_z\big)e_z(k) + \eta(k) + L_zv(k).$$

with $\mathbb{E}\{e_z(k)\} = 0$ for all $k \geq 0$, covariance matrix $\mathbb{E}\{e_z(k)e_z(k)^T|\{y(i)\}_{i=0}^{k-1}\} = \Sigma_z(k)$ and $\Sigma_z(0) \geq 0$.

The following theorem classifies all gains L_z of a non-optimal observer such that Problem 4.1 is solved.

Theorem 4.2. *For any $\Sigma_z(0) \geq 0$,*

$$\lim_{k\to\infty} \Sigma_z(k) = \tilde{\Sigma}_\infty = \begin{bmatrix} \tilde{P} & 0 \\ 0 & 0 \end{bmatrix},$$

if and only if $\rho(A_c) < 1$, $L_z = [L_1^T \ 0^T]^T$ and $L_1 \in \mathbb{R}^{n_x \times n_y}$ is chosen such that $\rho(A - L_1C) < 1$. Here, $\tilde{P}$ is the unique solution to

$$\tilde{P} = (A - L_1C)\tilde{P}(A - L_1C)^T + \Sigma_w + L_1\Sigma_vL_1^T,$$

and $\tilde{P} - P \geq 0$, where P is the unique solution to (4.7).

Proof. With $L_z = [L_1^T \ L_2^T]^T$ the error dynamics are

$$e_z(k+1) = \begin{bmatrix} A - L_1C & C_c \\ -L_2C & A_c \end{bmatrix} e_z(k) + \begin{bmatrix} w(k) - L_1v(k) \\ L_2v(k) \end{bmatrix}.$$

The error covariance matrix evolves as

$$\Sigma_z(k+1) = (A_z - L_z C_z)\Sigma_z(k)(A_z - L_z C_z)^T + \begin{bmatrix} \Sigma_w + L_1\Sigma_v L_1^T & L_1\Sigma_v L_2^T \\ L_2\Sigma_v L_1^T & L_2\Sigma_v L_2^T \end{bmatrix}. \tag{4.10}$$

Now we show that $\tilde{\Sigma}_\infty$ is the steady state solution of (4.10) if and only if $L_2 = 0$. First, we observe that if $L_2 = 0$ then $\tilde{\Sigma}_\infty$ is a steady state solution of (4.10), where $\tilde{P}$ is the solution to the Lyapunov equation

$$\tilde{P} = (A - L_1 C)\tilde{P}(A - L_1 C)^T + \Sigma_w + L_1\Sigma_v L_1^T.$$

Note that $\tilde{P} \geq 0$ exists and is unique if $\rho(A - L_1 C) < 1$. Second, if $\tilde{\Sigma}_\infty$ is a steady state solution of (4.10) the equations

$$\begin{aligned} \tilde{P} &= (A - L_1 C)\tilde{P}(A - L_1 C)^T + \Sigma_w + L_1\Sigma_v L_1^T, \\ 0 &= L_2(\Sigma_v L_1^T - C\tilde{P}(A - L_1 C)^T), \text{ and} \\ 0 &= L_2(C\tilde{P}C^T + \Sigma_v)L_2^T \end{aligned}$$

are fulfilled. The last equation is only fulfilled if $L_2 = 0$, since Σ_v is positive definite. This simultaneously fulfills the second equation. The first equation recovers the Lyapunov equation for $\tilde{P}$. Therefore, if $\tilde{\Sigma}_\infty$ is a steady state solution of (4.10) then $L_2 = 0$. Hence, (4.10) has $\tilde{\Sigma}_\infty$ as a steady state solution if and only if $L_2 = 0$.

Let us now look at the convergence of (4.10) to $\tilde{\Sigma}_\infty$. For any $\Sigma_z(0) \geq 0$, the error covariance matrix converges to $\tilde{\Sigma}_\infty$ if and only if $\rho(A_z - L_z C_z) < 1$. With $L_2 = 0$, the stability of $A_z - L_z C_z$ is guaranteed when both $\rho(A_c) < 1$ and $\rho(A - L_1 C) < 1$. Due to the detectability of (C, A) in Assumption 3.2 such a stabilizing L_1 exists. Therefore, (4.10) converges to $\tilde{\Sigma}_\infty$ for any $\Sigma_z(0) \geq 0$, if and only if $L_2 = 0$, $\rho(A - L_1 C) < 1$, and $\rho(A_c) < 1$. Further, $\rho(A_z - L_z C_z) < 1$ also makes $\tilde{\Sigma}_\infty$ the unique steady state solution of (4.10). Since the Kalman filter is the best linear estimator, we know that $\tilde{P} - P \geq 0$ and $\tilde{P} = P$ if $L_1 = APC^T(CPC^T + \Sigma_v)^{-1}$ [81]. This choice of L_1 turns the Lyapunov equation of $\tilde{P}$ into (4.7). □

Theorem 4.2 shows us that the attacker is able to use the non-optimal observer (4.9) solve to Problem 4.1, if and only if the controller is stable.

Corollary 4.2. *Problem 4.1 is solvable with a non-optimal observer of the form* (4.9) *if and only if* $\rho(A_c) < 1$.

According to Theorem 4.2, the attacker does not need to know the noise statistics Σ_w and Σ_v for the design of L_1 to estimate $x_c(k)$ perfectly, as long as L_1 is stabilizing. Hence, the attacker's required knowledge to solve Problem 4.1 is reduced when the operator uses a stable controller. Further, the attacker has a smaller computational burden when a time-invariant observer is used.

4.3 Defense Mechanisms

We presented under which conditions Problem 4.1 is solvable both with optimal and non-optimal strategies. Therefore, we investigate now how to prevent the attacker from estimating $x_c(k)$, i.e. make Problem 1 unsolvable. We present a defense mechanism and discuss why an unstable controller is only in certain cases a good defense mechanism.

4.3.1 Injecting Noise on the Controller Side

As previously shown, an attacker under Assumption 3.9 will be able to predict the controller state perfectly for $\rho(A_c) \leq 1$. We observe that the controller dynamics in (3.2) contain no uncertainty for the attacker when $y(k)$ is known. Therefore, an approach for defense is to introduce uncertainty in the form of an additional noise term on the controller side.

The additional noise term $\nu(k)$ has a zero mean Gaussian distribution with a positive semi-definite covariance matrix $\Sigma_\nu \in \mathbb{R}^{n_c \times n_c}$. Further, $\nu(k)$ is independent and identically distributed over time and also independent of $w(k)$, $v(k)$, and $z(0)$. The controller state with the additional noise term follows the dynamics

$$x_c(k+1) = A_c x_c(k) + B_c y(k) + \nu(k).$$

Here, $\nu(k)$ can be interpreted as process noise of the controller.

This changes the process noise of the closed-loop system (4.2) from $\eta(k)$ to $\tilde{\eta}(k) = [w(k)^T \ \nu(k)^T]^T$ such that

$$\mathbb{E}\left\{ \begin{bmatrix} \tilde{\eta}(k) \\ v(k) \end{bmatrix} \begin{bmatrix} \tilde{\eta}(k)^T & v(k)^T \end{bmatrix} \right\} = \left[\begin{array}{cc|c} \Sigma_w & 0 & 0 \\ 0 & \Sigma_\nu & 0 \\ \hline 0 & 0 & \Sigma_v \end{array}\right] = \left[\begin{array}{c|c} \tilde{Q} & 0 \\ \hline 0 & R \end{array}\right].$$

The following proposition shows that with $\nu(k)$, the attacker's desired covariance matrix Σ_∞ is not a steady state solution of (4.6) any more.

Proposition 4.3. The algebraic Riccati equation (4.6) with $Q = \tilde{Q}$ does *not* have Σ_∞ as a steady state solution.

Proof. With $\Sigma_z(k) = \Sigma_\infty$ and $Q = \tilde{Q}$ we obtain

$$A_z \Sigma_\infty A_z^T + \tilde{Q} = \begin{bmatrix} APA^T + \Sigma_w & 0 \\ 0 & \Sigma_\nu \end{bmatrix},$$

and using this in the Riccati equation (4.6) leads to

$$\Sigma_\infty = \begin{bmatrix} APA^T + \Sigma_w - APC^T(CPC^T + \Sigma_v)^{-1}CPA^T & 0 \\ 0 & \Sigma_\nu \end{bmatrix}.$$

For Σ_∞ to be a solution of (4.6) we need both

$$P = APA^T + \Sigma_w - APC^T(CPC^T + \Sigma_v)^{-1}CPA^T,$$

which, as shown previously, exists, and $\Sigma_\nu = 0$.

Since we assume $\Sigma_\nu \neq 0$, Σ_∞ is not a solution of (4.6) any more. □

Here, we see that the attacker will not be able to perfectly estimate the controller's state if we use this additional noise on the controller side even if the attacker knows the noise properties.

Remark 4.1. The approach of adding some additional noise to the system is quite similar to the watermarking approach used, for example, in [34]. The difference is that here the noise is added to the controller input, while in watermarking the noise is typically added to the output of the controller. Therefore, these results show that if we position the watermarking noise at a different position we get the additional benefit of the attacker not being able to estimate the state of the controller perfectly.

4.3.2 An Unstable Controller as Defense

As shown before, Problem 4.1 is not solvable if and only if $\rho(A_c) > 1$. Hence, designing the controller (A_c, B_c, C_c, D_c) such that $\rho(A'_z) < 1$ and $\rho(A_c) > 1$ leads to a successful defense against the discussed disclosure attack.

This implies that there are plants which have an inherent protection against the sensor attack. For example, all plants that are *not* strongly stabilizable, i.e. plants that cannot be stabilized with a stable controller [83], have an inherent protection against the estimation of the controller's state by the attacker. Further, there are also control strategies that give an inherent protection to the closed-loop system. Disturbance accommodation control [84], where the controller tries to estimate a persistent disturbance, is one example of these control strategies.

If a plant can be stabilized by using a stable controller, i.e. a strongly stabilizing plant, using an unstable controller instead comes with several issues. A fundamental limitation is that the integral of the sensitivity function is either zero for a stable open-loop system or equal to a constant value that depends on the unstable poles of the open-loop system and their directions for a multivariable discrete-time system [85]. As [86] shows with real world examples it can have dire consequences if this fundamental limitation is not taking into account properly. Hence, due to these fundamental limitation the introduction of unstable poles in the controller is not desirable. Another issue of unstable controllers is that an unstable controller leads to an unstable open-loop system, if the feedback loop is interrupted.

Therefore, using an unstable controller for a strongly stabilizing plant is not recommended, but is an appropriate defense mechanism if an unstable controller is needed to stabilize the plant.

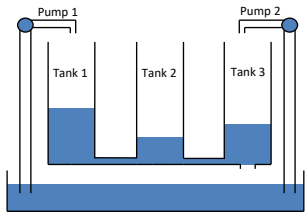

Figure 4.1: The three tank system

4.4 Simulations

In this section, we verify our results with simulations for a three tank system. After stating the model of the three tank system, we first show the effect of stable and unstable controllers on the attacker's estimate of the controller's state. Later, we verify that the additional noise prevents the attacker from estimating the controller's state perfectly.

4.4.1 The Three Tank System

For the simulation of the closed-loop system estimation by the attacker we look at the following continuous-time three tank system (see Figure 4.1)

$$\dot{x}(t) = \begin{bmatrix} -2 & 2 & 0 \\ 2 & -4 & 2 \\ 0 & 2 & -3 \end{bmatrix} x(t) + \begin{bmatrix} 0.5 & 0 \\ 0 & 0 \\ 0 & 0.5 \end{bmatrix} u(t) + w(t),$$
$$y(t) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} x(t) + v(t).$$

By discretizing the continuous-time system with a sampling period of $T_s = 0.5\,\mathrm{s}$ we obtain A, B, and C. We assume that $w(k) \sim \mathcal{N}(0, I_3)$ and $v(k) \sim \mathcal{N}(0, 0.1I_2)$.

4.4.2 Stable and Unstable Controllers

Now that the system matrices are defined we are going to verify that the controller's stability influences the estimates of the controller's state by the attacker. We consider an observer-based feedback controller

$$x_c(k+1) = (A - BK_i - LC)x_c(k) + Ly(k)$$
$$u(k) = -K_i x_c(k)$$

where L is the observer gain and K_i is the controller gain. The closed-loop system matrix is then

$$A'_{z,i} = \begin{bmatrix} A & -BK_i \\ LC & A - BK_i - LC \end{bmatrix}.$$

According to Assumption 3.3, $\rho(A'_{z,i}) < 1$, which means that K_i and L are designed such that $\rho(A - BK_i) < 1$ and $\rho(A - LC) < 1$. The matrix L is designed via pole placement to place the eigenvalues of $A - LC$ at 0.1, 0.2, and 0.3. Therefore, the error dynamics of the observer used in the controller are stable. In the following, we design three different K_i such that $\rho(A - BK_i) < 1$.

The first controller K_S places the poles of $A - BK_S$ at 0.4, 0.5, and 0.6. This first controller results in stable controller dynamics $A - BK_S - LC$ with $\rho(A - BK_S - LC) = 0.4167$.

The second controller, K_U, is unstable, i.e. $\rho(A - BK_U - LC) > 1$, but has no modes on the unit circle. We determine K_U, such that $\rho(A - BK_U) < 1$ and $A - BK_U - LC$ has an eigenvalue at 1.5. The controller we obtain is

$$K_U = \begin{bmatrix} 0.5530 & 1.9589 & 1.2225 \\ 1.8414 & 27.0785 & -12.9349 \end{bmatrix}$$

and it places the eigenvalues of $A - BK_U - LC$ at 1.5, -0.5175, and -0.1066 and the eigenvalues of $A - BK_U$ at 0.6275, $0.4272 + j0.6456$, and $0.4272 - j0.6456$.

For the design of the third controller, K_I, we place two controller eigenvalues inside the unit circle and one at 1, such that $\rho(A - BK_I - LC) = 1$, while guaranteeing that $\rho(A - BK_I) < 1$. We obtain

$$K_I = \begin{bmatrix} 3.0988 & -6.0472 & 2.3966 \\ 4.0471 & 10.8175 & -4.4516 \end{bmatrix},$$

which places the eigenvalues of $A - BK_I - LC$ at 1, -0.2227, and -0.3693 and the eigenvalues of $A - BK_I$ at -0.2669, $0.6405 + j0.5942$, and $0.6405 - j0.5942$.

For the first two controllers, the attacker designs a time-invariant Kalman filter with gain L_z^i and steady state error covariance matrix $\Sigma_\infty^i = \lim_{k\to\infty} \Sigma^i(k)$, where $i \in \{S, U\}$. The attacker's time-invariant Kalman filter design leads to an observer gain L_z^S for the closed-loop system, which matches our results in Theorem 4.2. Since K_U leads to an unstable controller, we know according to Corollary 4.2 that no time-invariant observer exists that solves Problem 4.1. Further, Corollary 4.1 shows that even if the attacker would use a time-varying Kalman filter, Problem 4.1 is not solvable.

For the closed-loop system with K_I, the attacker needs to use a time-varying Kalman filter to obtain a perfect estimate of $x_c(k)$. The error covariance matrix in this case will converge to the same as in the case with K_S.

Now that we designed the Kalman filters for each of the three closed-loop systems, let us look at the estimation error $e_z(k) = z(k) - \hat{z}(k) \in \mathbb{R}^6$. Here, we are only

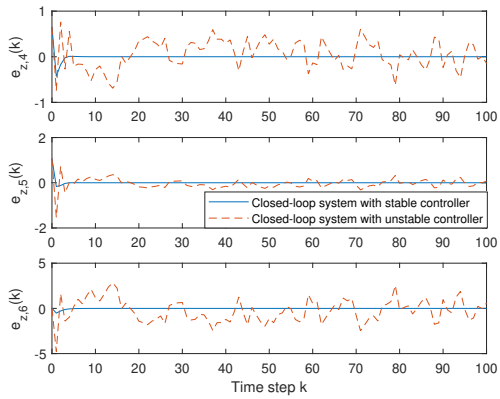

Figure 4.2: Comparison of the estimation error trajectories for the stable and unstable controller, K_S and K_U respectively

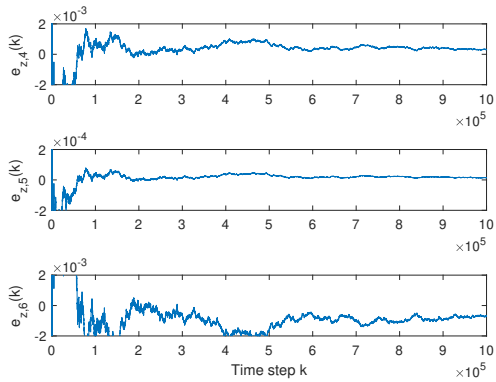

Figure 4.3: Estimation error of the controller's state when the controller has a pole on the unit circle and the attacker uses a time-varying Kalman filter

interest in the last three elements of $e_z(k)$, because they represent the estimation error of the controller state. The jth element of $e_z(k)$ is denoted by $e_{z,j}(k)$, where $j \in \{1, \cdots, 6\}$. Figure 4.2 shows that in case of a stable controller K_S the estimation error converges quickly to zero and the attacker obtains a perfect estimate of the controller's state. However, if we use an unstable controller K_U the estimation error remains noisy and the attacker is not able to obtain a perfect estimate of the controller's state. Furthermore, when K_I is used, we observe that the estimation error converges to zero, but is still not zero after a million time steps (see Figure 4.3). Theorem 4.1 only tells us that the error will converge, but we know it does not converge exponentially by Proposition 4.2. Although the attacker can obtain an almost perfect estimate with the time-varying Kalman filter after a million time

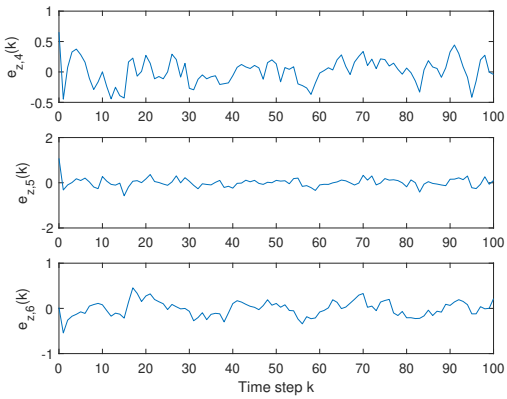

Figure 4.4: The effect of the additional noise on the estimation error of the controller's state when a stable controller is used

steps, it is still not a perfect estimate. This shows us that a controller with modes on the unit circle can prevent the attacker from quickly obtaining a perfect estimate.

4.4.3 Injecting Process Noise for the Controller

Now that we showed how the controller design affects the attacker's estimate of the controller's state, we verify that injecting noise to the input of the controller prevents the attacker from estimating $x_c(k)$ perfectly. The additional noise $\nu(k)$ has a covariance of $\Sigma_\nu = 0.01 I_3$. Since the attacker has full model knowledge, we assume that the attacker knows Σ_ν and designs its observer appropriately.

Figure 4.4 shows the trajectory of the estimation error of the controller's state in this case, when the operator uses the stable controller K_S and the attacker uses again a time-invariant Kalman filter. Compared to Figure 4.2, the estimation error exhibits noisy behavior and the attacker is not able to obtain a perfect estimate even though we use the stable controller K_S. Hence, the additional noise prevents the attacker from estimating the controller's state perfectly.

4.5 Summary

In this chapter, we investigated under which conditions an attacker according to Assumption 3.9 is able to estimate the controller's state perfectly. Although it may seem obvious that an attacker according to our attack model can always estimate the controller's state, we gave necessary and sufficient conditions when an attacker is not able to obtain a perfect estimate. These conditions state that unstable controller dynamics prevent the attacker from obtaining a perfect estimate. Further, the attacker can use a non-optimal time-invariant observer to perfectly estimate the controller state if and only if the controller has stable dynamics. Hence, if a

controller has eigenvalues inside or on the unit circle, the attacker is able to launch the attack in (3.11) after estimating the controller's state.

A defense mechanisms has been proposed to make the controller states confidential. This mechanism prevents the attacker from obtaining a perfect estimate by adding uncertainty to the controller dynamics. Furthermore, we discussed the use of an unstable controller, which gives an inherent protection to plants that are not strongly stabilizable. However, designing such a controller introduces fundamental limitations on the sensitivity function of the closed-loop system and should only be used when an unstable controller is needed to stabilize the plant. In the following chapter, we assume that the attack is executable and investigate the confidentiality of the anomaly detector state. With knowledge of the anomaly detector's state the attacker is able to design more powerful attacks.

Chapter 5

On the Confidentiality of the Detector State

5.1 Problem Formulation

In Chapter 4, we showed under which conditions the attacker is able to perfectly estimate $x_c(k)$ and also discussed possible defense mechanisms. In this chapter, we assume that the attacker has successfully obtained a perfect estimate of $x_c(k)$. Recall that with access to $x_c(k)$ the attacker has access to $\hat{x}(k)$. Hence, the attacker can execute the worst-case attack (3.11) and the detector dynamics become

$$\begin{aligned} x_D(k+1) &= \theta\big(x_D(k), a(k)\big), \\ y_D(k+1) &= d\big(x_D(k), a(k)\big). \end{aligned}$$

But without knowledge of $x_D(k)$, the attacker needs to design $a(k)$ conservatively to remain undetected, i.e. guarantee that $y_D(k+1) \leq J_D$ for all k (see Chapter 6). If the attacker wants to maximizes its attack potential, it needs to know $x_D(k)$.

Therefore, with full control over the detector's input, the attacker can try to generate an attack, which simultaneously remains undetected and helps with the estimation of $x_D(k)$. For simplicity we assume that the estimation of $x_D(k)$ starts at $k = 0$, such that $r(k) = a(k)$ for all $k \geq 0$ and $x_D(0)$ is unknown to the attacker. Moreover, in this chapter, we only investigate anomaly detectors with linear dynamics, i.e.

$$\begin{aligned} x_D(k+1) &= A_D x_D(k) + B_D r(k), \\ y_D(k+1) &= f_D\big(A_D x_D(k) + B_D r(k)\big) = d\big(x_D(k), r(k)\big). \end{aligned} \tag{5.1}$$

Here, $A_D \in \mathbb{R}^{n_D \times n_D}$, $B_D \in \mathbb{R}^{n_D \times n_y}$ has full rank, and $n_D \leq n_y$.
Further,

(i) $f_D(\cdot)$ is a vector norm on $\mathbb{R}^{n_D}$;

(ii) A_D needs to be Schur stable, i.e. $\rho(A_D) < 1$;

(iii) $g_D(A_D) < 1$, where $g_D(\cdot)$ the matrix norm on $\mathbb{R}^{n_D \times n_D}$ induced by $f_D(\cdot)$.

Hence, the first four detector conditions in Assumption 3.7 are fulfilled. Since the dynamics are linear we can, without loss of generality, rewrite the detector state as the superposition of two subsystems, i.e. $x_D(k) = x_{D,r}(k) + x_{D,a}(k)$, where

$$\begin{aligned} x_{D,a}(k+1) &= A_D x_{D,a}(k) + B_D a(k), \\ x_{D,r}(k+1) &= A_D x_{D,r}(k), \end{aligned}$$

with $x_{D,r}(0) = x_D(0)$ and $x_{D,a}(0) = 0$. Here, $x_{D,a}(k)$ is governed by the attack signal, while $x_{D,r}(k)$ is an autonomous system, which is governed by the initial state of the detector. Since A_D is Schur stable, $x_{D,r}(k)$ converges to zero as $k \to \infty$. This means that $x_{D,a}(k)$ can be seen as the estimate of $x_D(k)$ at time step k and $x_{D,a}(k)$ converges to $x_D(k)$ as $k \to \infty$.

To have a good estimate, i.e. reduce the uncertainty, at time step N, the attacker wants

$$||x_D(N) - x_{D,a}(N)||_2 = ||x_{D,r}(N)||_2 = \left|\left|A_D^N x_D(0)\right|\right|_2 \leq \gamma,$$

where $\gamma > 0$ is close to zero. Since $x_D(0)$ is unknown, we obtain an upper bound on $x_D(0)$

$$y_{\text{up}} = \max_x ||x||_2 \text{ subject to } x \in \{y \in \mathbb{R}^{n_y} : f_D(y) \leq J_D\}.$$

Based on y_{up}, we choose N such that

$$\left|\left|A_D^N\right|\right|_2 = \sigma_{\max}\big(A_D^N\big) \leq \frac{\gamma}{y_{\text{up}}}, \tag{5.2}$$

holds.

Remark 5.1. The slower $\sigma_{\max}(A_D^k)$ approaches zero as k grows large, the more time it takes for the attacker to obtain an accurate estimate of the $x_D(k)$. Hence, a defender can consider this fact, when designing the detector.

The attacker not only wants to reduce its uncertainty about $x_D(k)$, but also wants to remain undetected by the detector. If the detector triggers an alarm the operator will investigate it and might discover the attacker. This could lead to countermeasures against the attack. Therefore, we look now at the condition for the attacker to remain undetected. Since $f_D(\cdot)$ is a vector norm, we can determine the following condition to avoid detection.

$$\begin{aligned} y_D(k) &= f_D\big(x_{D,r}(k) + x_{D,a}(k)\big) \\ &\leq f_D\big(x_{D,r}(k)\big) + f_D\big(x_{D,a}(k)\big) \\ &\leq g_D\big(A_D^k\big) f_D\big(x_{D,r}(0)\big) + f_D\big(x_{D,a}(k)\big) \\ &\leq g_D\big(A_D^k\big) J_D + f_D\big(x_{D,a}(k)\big) \leq J_D, \\ \Rightarrow y_{D,a}(k) &= f_D\big(x_{D,a}(k)\big) \leq J(k), \end{aligned}$$

where $J(k) = J_D - g_D\big(A_D^k\big)J_D > 0$ for all $k > 0$. We see that if $y_{D,a}(k) \leq J(k)$, then the attack remains undetected. Note that $J(k)$ approaches J_D as $k \to \infty$. We can interpret $x_{D,a}(k)$ and $y_{D,a}(k)$ as a *virtual detector* with threshold $J(k)$ that the attacker initializes at $x_{D,a}(0) = 0$ and uses it to design its stealthy attack.

Let us summarize these results in a proposition and then discuss how the attacker can design $\{a(k)\}_{k=0}^{N-1}$.

Proposition 5.1. An attacker can reconstruct the detector state with accuracy γ, in N time steps, where N is such that $\sigma_{\max}\big(A_D^N\big) \leq \frac{\gamma}{y_{\text{up}}}$ is fulfilled. Furthermore, the attacker can simultaneously inject $a(k)$ satisfying $y_{D,a}(k+1) = d\big(x_{D,a}(k), a(k)\big) \leq J(k+1)$ for all $k \geq 0$ to remain undetected.

A simple way for the attacker to choose the detector input is $a(k) = 0$ for $k \in \{0, \cdots, N-1\}$. Then $x_{D,a}(k) = 0$ for all $k \geq 0$, which implies that $y_{D,a}(k) = 0 \leq J(k)$ for all $k \geq 0$. However, this leads to suspicious behavior in $y_D(k+1)$, for example an exponential decay of $y_D(k+1)$, which might raise an operator's suspicion when seeing this on the display in the control center. Another way is to not change the measurements and just observe $r(k)$ and feed it into $x_{D,a}(k)$. The advantage is that $y_D(k+1)$ behaves exactly as in the nominal case, but without knowledge of $x_D(k)$ any $r(k)$ might lead to an alarm, which is considered as a false alarm under nominal conditions. A third option is to make the alarm look like a false alarm, by inducing a spike in one element of $r(k)$, such that $x_D(k)$ is reset to zero. However, since the attacker is present in the system, this "false alarm" in the last two strategies might lead to the detection of the attacker if the operator decides to investigate the alarms. Figure 5.1 shows the detector output for the first two cases, when a MEWMA detector is used. Since the detector output for the third case would have a similar trajectory as the one for the second case, it is not displayed in Figure 5.1.

Therefore, to make sure to remain undetected and not raise the operator's suspicion an attacker needs to design $a(k)$ appropriately. Since $a(k)$ has a direct influence on $y_D(k+1)$, the attacker tries to design $a(k)$ in such a way that under attack $y_D(k+1)$ approximately has probability density $q_{k+1}(y_D)$, but no alarms are caused. Since $x_{D,a}(k)$ converges to $x_D(k)$ as $k \to \infty$, we also get that $y_{D,a}(k)$ converges to $y_D(k)$ as $k \to \infty$. Therefore, we look at the virtual detector $y_{D,a}(k)$ instead of $y_D(k)$, since the attacker has no direct access to $y_D(k)$. This means when $y_D(k)$ has a probability density function $q_k(y_D)$, which changes for a given $x_D(k-1)$ then we assume that $y_{D,a}(k)$ has the same probability density function $q_k(y_D)$ but given $x_{D,a}(k-1)$. The first problem is to find a probability density function $p_k(y_D)$ that approximates $q_k(y_D)$.

Problem 5.1. Find $p_{k+1}(y_D)$, such that $\text{supp}\big(p_{k+1}(y_D)\big) = [0, J(k+1)]$ (no alarms) and $p_{k+1}(y_D)$ resembles $q_{k+1}(y_D)$ as closely as possible to not raise the suspicion of the operator.

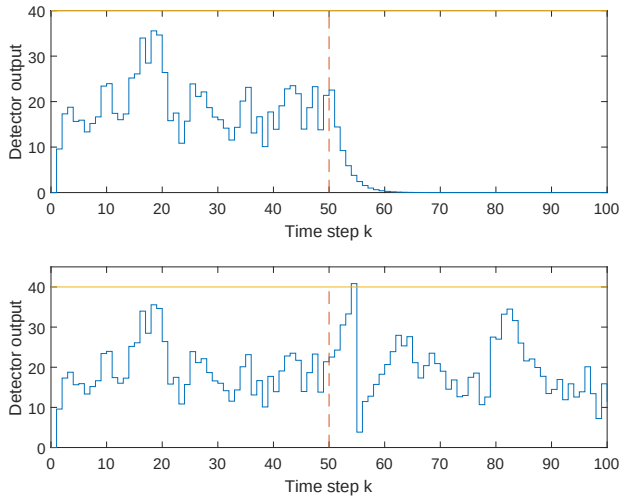

Figure 5.1: At time step 50 the attacker gains access to the detector input, i.e. $r(k) = a(k)$. The upper plot shows that choosing $a(k) = 0$ leads to a suspicious exponential decay of the detector output. The lower plot shows that not changing the detector input, i.e. $a(k) = r(k)$ for $k \geq 50$, might trigger an alarm. Both plots could lead to the detection of the attack, either by raising the suspicion of the operator or by triggering an alarm that is then investigated.

After $p_{k+1}(y_D)$ is found, the attacker can try to design $a(k)$, such that $y_{D,a}(k+1)$ follows samples from $p_{k+1}(y_D)$. Further, we want to investigate if the attacker can not only design $a(k)$ such that the distribution of $y_D(k+1)$ is $p_{k+1}(y_D)$ but also such that it can have an impact on the plant. More specifically, the attacker wants to maximize the maximum average estimation error of the critical states $||e_{crit}(k)||_\infty$. Here, $e_{crit}(k) = \mathbb{E}\{T_{crit}e(k)\}$, where $T_{crit} \in \mathbb{R}^{n_{crit} \times n_x}$ is a matrix that extracts the critical estimation errors of $e(k)$ and $n_{crit} \leq n_x$. This could, for example, be the estimation error of the pressure in a closed container, which might explode if the pressure is too large.

Problem 5.2. Design $a(k)$ such that $y_{D,a}(k+1) = s_{k+1}$ and $||e_{crit}(k+1)||_\infty$ is maximized, where s_{k+1} is a sample from the probability distribution with probability density function $p_{k+1}(y_D)$.

To estimate $x_D(k)$ and remain undetected the attacker needs to solve Problem 5.1 and Problem 5.2. In the following, we show how an attacker can solve these problems.

5.2 Problem 5.1: How to Characterize $p_k(y_D)$

Kullback *et al.* [87] defined the average information gain of each observation to distinguish between a hypothesis with density function $p(y_D)$ and a hypothesis with density function $q(y_D)$ as $D_{KL}(p||q) = \int p(y_D) \ln\left(\frac{p(y_D)}{q(y_D)}\right) dy_D$, which is known as the Kullback-Leibler divergence. Furthermore, $D_{KL}(p||q)$ is convex in the pair of its arguments. Therefore, it comes quite natural to minimize the average information gain $D_{KL}(p_k||q_k)$ to find $p_k(y_D)$, such that $p_k(y_D)$ becomes hard to distinguish from $q_k(y_D)$. The optimization problem is

$$\begin{aligned} &\min_{p_k(y_D)} \int_0^{J(k)} p_k(y_D) \ln\left(\frac{p_k(y_D)}{q_k(y_D)}\right) dy_D \\ &\text{s.t.} \begin{cases} p_k(y_D) \geq 0 & \forall y_D \in [0, J(k)] \\ p_k(y_D) = 0 & \forall y_D \notin [0, J(k)] \\ \int_0^{J(k)} p_k(y_D) dy_D = 1 \\ \text{more convex constraints on } p_k(y_D) \end{cases} . \end{aligned} \tag{5.3}$$

The first three constraints are necessary such that $p_k(y_D)$ is a probability density function. One can also impose more convex constraints on $p_k(y_D)$, which preserve the convexity of the problem. For example, we can impose a constraint on the mean $\int_0^{J(k)} y_D p_k(y_D) dy_D$ or the second raw moment $\int_0^{J(k)} y_D^2 p_k(y_D) dy_D$ as well.

Here, we only look at the case where no additional constraints are imposed. Then, we need to solve

$$\begin{aligned} &\min_{p_k(y_D)} \int_0^{J(k)} p_k(y_D) \ln\left(\frac{p_k(y_D)}{q_k(y_D)}\right) dy_D \\ &\text{s.t.} \begin{cases} p_k(y_D) \geq 0 & \forall y_D \in [0, J(k)] \\ p_k(y_D) = 0 & \forall y_D \notin [0, J(k)] \\ \int_0^{J(k)} p_k(y_D) dy_D = 1 \end{cases} . \end{aligned} \tag{5.4}$$

It turns out that the solution to (5.4) is the truncated version of $q_k(y_D)$.

Proposition 5.2. The optimizer to (5.4) is

$$p_k^*(y_D) = \begin{cases} \frac{q_k(y_D)}{\int_0^{J(k)} q_k(y_D) dy_D} & y_D \in [0, J(k)] \\ 0 & \text{otherwise} \end{cases}, \tag{5.5}$$

i.e. the truncated version of $q_k(y_D)$ is the optimal solution.

Proof. Let $\lambda \in \mathbb{R}$ be a Lagrange multiplier and the Lagrangian be

$$\begin{aligned} L(p,\lambda) &= \int_0^{J(k)} p_k(y_D) \ln\left(\frac{p_k(y_D)}{q_k(y_D)}\right) dy_D + \lambda\left(\int_0^{J(k)} p_k(y_D) dy_D - 1\right) \\ &= \int_0^{J(k)} p_k(y_D) \ln\left(\frac{p_k(y_D)}{q_k(y_D)}\right) + \lambda\left(p_k(y_D) - \frac{1}{J(k)}\right) dy_D \\ &= \int_0^{J(k)} l\big(p_k(y_D),\lambda\big) dy_D. \end{aligned}$$

A necessary condition for optimality (see [88]) is

$$\frac{d}{dp_k(y_D)} l\big(p_k(y_D),\lambda\big)\Big|_{p_k(y_D)=p_k^*(y_D)} = 0.$$

Solving for $p_k^*(y_D)$ leads to

$$p_k^*(y_D) = \begin{cases} e^{-1-\lambda} q_k(y_D) & \forall y_D \in [0, J(k)] \\ 0 & \forall y_D \notin [0, J(k)] \end{cases},$$

where we already incorporated the first two constraints of (5.4). Now we use the last constraint to find

$$\lambda = -1 + \ln\left(\int_0^{J(k)} q_k(y_D) dy_D\right),$$

which results in $p_k^*(y_D)$. □

5.3 Problem 5.2: How to Characterize $a(k)$

Once we determined $p_{k+1}(y_D)$, we take a sample from this distribution. Let the obtained sample be s_{k+1}. Now we want to design $a(k)$ such that

$$y_{D,a}(k+1) = f_D\big(A_D x_{D,a}(k) + B_D a(k)\big) = s_{k+1}.$$

As mentioned before the attacker also wants to maximize the operator's average estimation error of the critical system states. From Chapter 3, we know that the average value of the error dynamics under attack evolve as

$$e_a(k+1) = A e_a(k) - L\Sigma_r^{\frac{1}{2}} a(k), \tag{5.6}$$

with $e_a(0) = 0$. Therefore the average estimation error of critical states is $e_{crit}(k) = T_{crit} e_a(k)$. The optimization problem to find $a(k)$ becomes then

$$\begin{aligned} \mathcal{I}_e := \max_{a(k)} ||T_{crit} e_a(k+1)||_\infty &= \max_{a(k)} \big|\big|T_{crit} A e_a(k) - T_{crit} L\Sigma_r^{\frac{1}{2}} a(k)\big|\big|_\infty \\ \text{s.t. } y_{D,a}(k+1) &= f_D\big(A_D x_{D,a}(k) + B_D a(k)\big) = s_{k+1}, \end{aligned} \tag{5.7}$$

where both $e_a(k)$, $x_{D,a}(k)$, and s_{k+1} are known to the attacker.

Before we introduce the solution to (5.7), we define the dual norm of a vector norm [89].

Definition 5.1. The dual norm of a vector norm $f_D(x)$ in $\mathbb{R}^n$ is defined as

$$f^{\mathcal{D}}(z) := \max_x \left|z^T x\right| \text{ s.t. } f_D(x) = 1,$$

where z is a vector in $\mathbb{R}^n$.

Now we introduce an intermediate result for solving (5.7).

Lemma 5.1. *The optimal value $\mathcal{I}$ of*

$$\max_{\bar{a}} |\bar{c}^T \bar{a} + \bar{d}| \text{ s.t. } f_D(\bar{a}) = s, \tag{5.8}$$

where $s \geq 0$, $\bar{d} \in \mathbb{R}$, $\bar{a}, \bar{c} \in \mathbb{R}^{n_D}$, is given by

$$\mathcal{I} = \max\left(|f^{\mathcal{D}}(\bar{c})s + \bar{d}|, |-f^{\mathcal{D}}(\bar{c})s + \bar{d}|\right) \tag{5.9}$$

with the maximizer

$$\bar{a}^* = \arg\max_{\bar{a}} (-1)^j c^T \bar{a} \text{ s.t. } f_D(\bar{a}) \leq s. \tag{5.10}$$

Here, $j = 2$ if $\left|f^{\mathcal{D}}(\bar{c})s + \bar{d}\right| \geq \left|-f^{\mathcal{D}}(\bar{c})s + \bar{d}\right|$ and $j = 1$ otherwise.

Proof. We first split (5.8) into two optimization problems, one that maximizes and one that minimizes $\bar{c}^T \bar{a} + \bar{d}$ under the given constraint, respectively. The larger absolute value of the optimal values of these two problems gives us the solution to (5.8). Note that $\bar{d}$ is a scalar and, therefore, the optimizer of these two problems will maximize or minimize $\bar{c}^T \bar{a}$, respectively. Definition 5.1 gives us that $\max_{f_D(\bar{a})=s} \left|\bar{c}^T \bar{a}\right| = f^{\mathcal{D}}(\bar{c})s$, from which (5.9) readily follows. Since the optimizer lies on the boundary of the constraint set, we replace the equality constraint of (5.8) with an inequality constraint to obtain the convex optimization (5.10). □

Before we present the main result of this section, let us introduce t_i^T as the ith row of T_{crit}, $\bar{c}_i^T = -t_i^T L \Sigma_r^{\frac{1}{2}} B_D^\dagger$, and $\bar{d}_i = t_i^T \left(Ae_a(k) + L\Sigma_r^{\frac{1}{2}} B_D^\dagger A_D x_{D,a}(k)\right)$.

Theorem 5.1. *The solution $\mathcal{I}_e$ of (5.7) is given by*

$$\mathcal{I}_e = \max_{i \in \{1, \cdots, n_{crit}\}} \max\left(\left|f^{\mathcal{D}}(\bar{c}_i)s_{k+1} + \bar{d}_i\right|, \left|-f^{\mathcal{D}}(\bar{c}_i)s_{k+1} + \bar{d}_i\right|\right),$$

and the corresponding attack vector can be found as

$$a(k) = B_D^\dagger \left(\bar{a}^* - A_D x_{D,a}(k)\right) \tag{5.11}$$

with $\bar{a}^*$ being the optimizer of the convex problem

$$\bar{a}^* = \arg\max_{\bar{a}} (-1)^{j_{i^*}} \bar{c}_{i^*}^T \bar{a} \text{ s.t. } f_D(\bar{a}) \leq s_{k+1}$$

where $\bar{a} \in \mathbb{R}^{n_D}$, $i^* \in \{1, \cdots, n_{crit}\}$ denotes an element of $T_{crit} e_a(k+1)$ for which $\mathcal{I}_e$ is achieved, and $j_{i^*} = 2$ if $\left| f^{\mathcal{D}}(\bar{c}_{i^*}) s_{k+1} + \bar{d}_{i^*} \right| \geq \left| -f^{\mathcal{D}}(\bar{c}_{i^*}) s_{k+1} + \bar{d}_{i^*} \right|$ and $j_{i^*} = 1$ otherwise.

Proof. Recall from Chapter 3, that we can write

$$||T_{crit} e_a(k+1)||_\infty = \max_{i \in \{1, \cdots, n_{crit}\}} \left| t_i^T e_a(k+1) \right|,$$

where $t_i^T e_a(k+1)$ represents the estimation error of the ith critical state. Therefore, we can solve n_{crit} problems of the form

$$\begin{aligned} \max_{a(k)} & \left| t_i^T A e_a(k) - t_i^T L \Sigma_r^{\frac{1}{2}} a(k) \right| \\ & \text{s.t. } f\big(A_D x_{D,a}(k) + B_D a(k)\big) = s_{k+1}, \end{aligned} \tag{5.12}$$

where $i \in \{1, \cdots, n_{crit}\}$ and pick $a(k)$ which results in the maximal objective value of all of these problems. Introducing $\bar{a} = A_D x_{D,a}(k) + B_D a(k)$, we reformulate (5.12) as

$$\max_{\bar{a}} \left| \bar{c}_i^T \bar{a} + \bar{d}_i \right| \text{ s.t. } f_D(\bar{a}) = s_{k+1}, \tag{5.13}$$

which represents n_{crit} problems of the form presented in Lemma 5.1. Therefore, we can use Lemma 5.1 to determine both $\mathcal{I}_e$, and $\bar{a}$ and with that $a(k)$. □

Theorem 5.1 shows us that, while estimating $x_D(k)$, the attacker can simultaneously design $a(k)$ such that the attack maximizes the estimation error at each time step.

Remark 5.2. If $f_D(x) = \left(\sum_i |x_i|^p \right)^{\frac{1}{p}}$, where $1 \leq p \leq \infty$, and x_i is the ith element of x, then $f^{\mathcal{D}}(x) = \left(\sum_i |x_i|^q \right)^{\frac{1}{q}}$ such that $\frac{1}{p} + \frac{1}{q} = 1$. This is a result of the Hölder inequality (see [89]).

Remark 5.3. One can also think of solutions that take other objectives into account when designing $a(k)$ at each time step. However, we choose this objective, because it maximizes the estimation error of the critical state in the sense of the maximum norm and we are able to find an analytical solution.

5.4 Application to the MEWMA Detector

Now we apply the previously presented procedure to the MEWMA detector and give an illustrative example of the control of a wind-excited tall building.

Recall the MEWMA detector is defined as

$$\begin{aligned} x_D(k+1) &= \beta r(k) + (1-\beta)x_D(k), \\ \tilde{y}_D(k+1) &= \frac{2-\beta}{\beta}||x_D(k+1)||_2^2, \end{aligned} \tag{5.14}$$

where $\beta \in (0,1]$. If $\tilde{y}_D(k+1) \leq \tilde{J}_D$ no alarm is triggered, where $\tilde{J}_D \in \mathbb{R}_{\geq 0}$ is a predefined threshold. Otherwise, an alarm is triggered and the detector state is reset to zero. The MEWMA detector as in (5.14), does not fit the detector model in (5.1), but we can rewrite it as

$$\begin{aligned} x_D(k+1) &= \beta r(k) + (1-\beta)x_D(k), \\ y_D(k+1) &= ||x_D(k+1)||_2, \end{aligned} \tag{5.15}$$

and use $J_D = \sqrt{\frac{\beta}{2-\beta}\tilde{J}_D}$ as the new detector threshold. This now fits (5.1) with $A_D = (1-\beta)$, $B_D = \beta$, $f_D(\cdot)$ being the Euclidean norm, and $g_D(\cdot) = \sigma_{\max}(\cdot)$.

Recall, $x_D(0)$ is unknown to the attacker. Since the dynamics are linear, we write the MEWMA detector as the superposition of two subsystems, $x_{D,a}(k)$ and $x_{D,r}(k)$, so that $x_D(k) = x_{D,a}(k) + x_{D,r}(k)$, where

$$\begin{aligned} x_{D,a}(k+1) &= \beta a(k) + (1-\beta)x_{D,a}(k), \\ x_{D,r}(k+1) &= (1-\beta)x_{D,r}(k), \end{aligned}$$

$k \geq 0$, $x_{D,r}(0) = x_D(0)$, and $x_{D,a}(0) = 0$.

Now we determine the attack duration N according to (5.2).

Proposition 5.3. The uncertainty of the MEWMA detector's state at time step N is smaller than $\gamma > 0$, i.e. $||x_D(N) - x_{D,a}(N)||_2 \leq \gamma$ if

$$N \geq \left\lceil \frac{\ln(\frac{\gamma}{J_D})}{\ln(1-\beta)} \right\rceil. \tag{5.16}$$

Proof. Since $A_D = 1-\beta$, we see that $\sigma_{\max}(A_D^N) = (1-\beta)^N$. Further, we determine that $y_{up} = J_D$. With that we solve (5.2) for N and obtain inequality (5.16). □

The attacker can launch an attack for N time steps such that the initial detector state $x_D(0)$ decreased sufficiently. This means that the attacker's uncertainty about $x_D(k)$ at time step N is small, i.e. $x_D(N) \approx x_{D,a}(N)$. Note that for $N \geq 0$ we need $\gamma \leq J_D$. Further, for the attack to remain undetected we obtain $J(k) = J_D(1-(1-\beta)^k)$, because $g_D(A_D^k) = (1-\beta)^k$.

Now that we have determined N and $J(k)$ let us derive the probability density function $p_k(y_D)$ by finding $q_k(y_D)$ under nominal conditions. Here, we change the procedure of this chapter slightly and look at

$$\frac{1}{\beta^2}y_D(k+1)^2 = ||r(k) + \frac{1-\beta}{\beta}x_D(k)||_2^2, \tag{5.17}$$

instead of $y_D(k+1)$. In the nominal case, (5.17) follows a noncentral χ^2 distribution with n_y degrees of freedom and noncentrality parameter $\lambda(k+1) = \left(\frac{1-\beta}{\beta}\right)^2 x_D(k)^T x_D(k)$ at each time step.

Therefore, according to Proposition 5.2, we design $p_{k+1}(y_D)$ as a truncated noncentral χ^2 distribution with n_y degrees of freedom, noncentrality parameter $\lambda_a(k+1) = \left(\frac{1-\beta}{\beta}\right)^2 x_{D,a}(k)^T x_{D,a}(k)$ and support $\text{supp}(p_{k+1}(y_D)) = \left[0, \frac{1}{\beta^2} J(k+1)^2\right]$.

After we draw a sample s_{k+1} from the truncated noncentral χ^2 distribution $p_{k+1}(y_D)$, we use (5.7) to determine $a(k)$, which for the MEWMA case looks as follows

$$\mathcal{I}_e = \max_{a(k)} ||T_{crit} e_a(k+1)||_\infty \text{ s.t. } ||\beta a(k) + (1-\beta) x_{D,a}(k)||_2 = \beta\sqrt{s_{k+1}}. \quad (5.18)$$

We can then directly use Theorem 5.1 to derive the impact for the MEWMA detector. Let us we first introduce $\bar{c}_i^T = -\frac{1}{\beta} t_i^T L \Sigma_r^{\frac{1}{2}}$ and $\bar{d}_i = t_i^T \left(A e_a(k) + \frac{1-\beta}{\beta} L \Sigma_r^{\frac{1}{2}} x_{D,a}(k)\right)$.

Corollary 5.1. *The impact for the MEWMA detector is*

$$\mathcal{I}_e^M = \max_{i \in \{1,\cdots,n_c\}} \max\left(\left|||\bar{c}_i||_2 \beta\sqrt{s_{k+1}} + \bar{d}_i\right|, \left| - ||\bar{c}_i||_2 \beta\sqrt{s_{k+1}} + \bar{d}_i\right|\right)$$

for the attack vector

$$a(k) = (-1)^{j_{i^*}} \frac{\bar{c}_{i^*}}{||\bar{c}_{i^*}||_2}\sqrt{s_{k+1}} - \frac{1-\beta}{\beta} x_{D,a}(k),$$

where i^* is an index that results in $\mathcal{I}_e^M$, and

$$j_{i^*} = \begin{cases} 2 & \text{if } \left|||\bar{c}_i||_2 \beta\sqrt{s_{k+1}} + \bar{d}_i\right| \geq \left| - ||\bar{c}_i||_2 \beta\sqrt{s_{k+1}} + \bar{d}_i\right|, \\ 1 & \text{otherwise.} \end{cases}$$

5.4.1 Illustrative Example

To verify the procedure for the MEWMA detector, we investigate the example of the excitation of tall buildings by wind. Figure 5.2 illustrates the effect wind can have on a tall building. Yang *et al.* [90] give a benchmark problem for the active control of a wind-excited building. For the simulation we use the linearized twelve dimensional reduced order model of a 76 story building given in [90]. We use the model with $n_y = 20$ measurements and further discretize it with a sampling period of $T_s = 0.01\,\text{s}$ to design the linear-quadratic-Gaussian controller. For the MEWMA detector, we use $\beta = 0.2$ and $\tilde{J}_D = 40$. To reduce the uncertainty about $x_D(k)$, we choose $\gamma = 10^{-6}$ such that we get an attack length of $N = 66$ time steps according to (5.16). We let the system run for 100 time steps initially and then start the attack at $k = 100$, to obtain a comparison of $\tilde{y}_D(k)$ before and after the attack. Further, for this simulation we use $T_{crit} = I_{n_x}$.

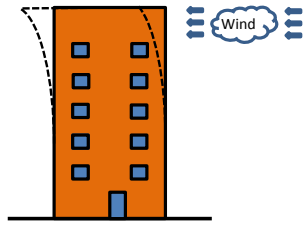

Figure 5.2: Illustration of a building moving under the influence of wind

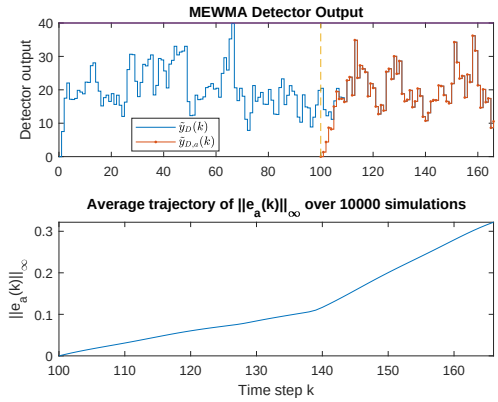

Figure 5.3: The upper plot shows how $\tilde{y}_D(k)$ behaves before and after the attack starting at $k = 100$. The lower plot shows the average trajectory of $||e_a(k)||_\infty$ over 10000 simulations

The upper plot in Figure 5.3 shows one simulation run of the trajectory of $\tilde{y}_D(k)$ for the MEWMA detector as described in (5.14) before and after the attack. The trajectory of output of the virtual MEWMA detector $\tilde{y}_{D,a}(k)$ is also displayed in that plot. We see that the trajectory $\tilde{y}_D(k)$ after the attack is still random and does not show any obvious irregularities to the bare human eye. Furthermore, the attack is not detected since the alarm threshold $\tilde{J}_D = 40$ is never crossed and we observe that $\tilde{y}_{D,a}(k)$ converges to $\tilde{y}_D(k)$ as time progresses. This shows us that the attacker's estimate of $x_D(k)$ becomes more accurate over time. Finally, we look at the accuracy of the estimate at the end of the attack. We have $||x_D(166) - x_{D,a}(166)||_2 = 6.0573 \cdot 10^{-7}$. As desired, the uncertainty is smaller than $\gamma = 10^{-6}$.

The lower plot of Figure 5.3 shows the average trajectory of $||e_a(k)||_\infty$ over 10000 simulations. Here, we see that the maximum estimation error is on average

increasing during the attack.

Therefore, we verified that an attacker with access to and control over the measurements is able to break the confidentiality of the internal state of the MEWMA detector and to simultaneously increase the maximum estimation error of the critical states.

5.5 Summary

In this chapter, we presented how an attacker can obtain an estimate of the internal state of an anomaly detector with linear dynamics. The attacker utilizes the detector dynamics to create a virtual detector that is used to design a stealthy attack. We use the Kullback-Leibler divergence to find a probability distribution that mimics the distribution of the nominal detector output. With this distribution the attacked detector output looks nominal and will not raise the operator's suspicion. Samples from this distributions are drawn to characterize the attack signal that simultaneously maximizes the average estimation error of critical plant states at each time step by exploiting the dual norm of a vector norm. We verify that this attack is working by applying it to a MEWMA detector. Together with Chapter 4, this chapter showed how an attacker can obtain the additional knowledge of the internal controller and detector states needed for the worst-case attack presented in Chapter 3. Therefore, the next chapters will compare the performance of different detectors in mitigating the attack impact and present methods to pick an optimal detector threshold.

Chapter 6

Comparison of Detectors

Chapters 4 and 5 showed that under certain conditions the attacker is able to gather the additional knowledge needed for the worst-case attack (3.11). In the first part of this chapter, we consider that the attacker is able to execute the worst-case attack (3.11). Since the attack impact will change with the detector used, we compare the performance of the χ^2, CUSUM, and MEWMA detectors using the metric proposed in [25] for two different processes.

In the second part of this chapter, we present a new metric to compare detectors, which depends neither on the attacker's objective nor on the system dynamics. This metric depends only on the number of sensors and the detector used. Recall that the detector dynamics under the worst-case attack (3.11) are

$$\begin{aligned} x_D(k+1) &= \theta\big(x_D(k), a(k)\big), \\ y_D(k+1) &= d\big(x_D(k), a(k)\big), \end{aligned}$$

where $x_D(k)$ is the detector state and $a(k)$ is the input of the detector designed by the attacker. To determine the new metric, we make use of a time-invariant set $\mathcal{B}$ for each detector, such that if $a(k) \in \mathcal{B}$ the attack is guaranteed to remain undetected independent of $x_D(k)$. Finally, we discuss the new metric and compare it with the results of the first part of this chapter.

6.1 Comparison of the χ^2, CUSUM, and MEWMA detectors

In this section, we use the metric of [25] to compare the performance of the χ^2, CUSUM, and MEWMA detectors. Recall that this metric plots the attack impact over the mean time between false alarms. Let us first introduce the way we determine the impact of the sensor attack.

6.1.1 Impact Estimation

Here, we consider that the operator uses the observer-based controller introduced in Section 3.4. Assuming the attacker executes the worst-case attack from time step $\underline{k}$ to $\bar{k}$, recall that the attack impact of Definition 3.2 is given by

$$\mathcal{I}(\tau) := \max_{a, x_D(\underline{k})} f(a) \quad \text{s.t.} \quad d\big(x_D(k), a(k)\big) \leq J_D \quad \forall\, k \in \Gamma,$$

where τ is the mean time between false alarms, $f(a)$ is the attacker's objective, $\Gamma \in [\underline{k}, \bar{k}]$, and $a = \{a(k)\}_{k=\underline{k}}^{\bar{k}}$ is the attack trajectory. Since $J_D = g(\tau)$ the impact depends on τ. Recall that, due to the linearity, we can write the closed-system as the superposition of two subsystems, where one system is excited by the noise and the initial condition, while the other system has zero initial condition and is excited by the attack. We only look at the subsystem that is excited by the worst-case attack,

$$\begin{bmatrix} x_a(k+1) \\ e_a(k+1) \end{bmatrix} = \underbrace{\begin{bmatrix} A - BK & BK \\ 0 & A \end{bmatrix}}_{=A_x} \begin{bmatrix} x_a(k) \\ e_a(k) \end{bmatrix} + \underbrace{\begin{bmatrix} 0 \\ -L\Sigma_r^{\frac{1}{2}} \end{bmatrix}}_{=B_x} a(k) \tag{6.1}$$

with $x_a(\underline{k}) = e_a(\underline{k}) = 0$. Since the attacked subsystem is linear,

$$\begin{aligned} x_a(\bar{k}) &= \begin{bmatrix} I_{n_x} & 0 \end{bmatrix} \left(\sum_{i=\underline{k}}^{\bar{k}} A_x^{\bar{k}-\underline{k}-i} B_x a(i) \right) \\ &:= T_{xa} a, \end{aligned} \tag{6.2}$$

where $T_{xa} \in \mathbb{R}^{n_x \times n_y(\bar{k}-\underline{k}+1)}$. We observe that if $a(k) = 0$ for all $k \in \Gamma$ the state of the attacked subsystem is zero as well. Therefore, we consider an attacker who wants to maximize $x_a(\bar{k})$. More specifically, the attacker wants to maximize $||x_a(\bar{k})||_\infty = ||T_{xa} a||_\infty$. Further, for the sake of simplicity, we assume the following for the rest of this chapter.

Assumption 6.1. The detector state at the beginning of the attack is zero, i.e. $x_D(\underline{k}) = 0$.

This leads to the following problem to determine the impact of the stealthy attack.

Problem 6.1. Find the global solution of

$$\max_{a} ||T_{xa} a||_\infty \quad \text{s.t.} \quad d\big(x_D(k), a(k)\big) \leq J_D \quad \forall\, k \in \Gamma. \tag{6.3}$$

to determine the worst-case impact of the sensor attack (3.11) on the closed-loop system (6.1) under the χ^2, CUSUM or MEWMA detectors, where T_{xa} is defined as in (6.2).

To find solution of Problem 6.1, we first show that the constraint set for a in (6.3) is a convex set.

Proposition 6.1. The constraint set for a,

$$d\big(x_D(k), a(k)\big) \leq J_D \quad \forall\, k \in \Gamma,$$

under Assumption 6.1 is non-empty and convex for the χ^2 detector, the CUSUM detector, and the MEWMA detector.

Proof. We begin by noting that $a = 0$ always fulfills the constraints. Hence, the constraint set is non-empty. To show the convexity of the set for the three detectors, we start by showing the convexity for the χ^2 detector. The constraints given by the χ^2 detector are

$$a(k)^T a(k) \leq J_D^{\chi^2} \quad \forall k \in \Gamma.$$

Since $a(k)^T a(k)$ represents a convex function for each $k \in \Gamma$, we know that the union of the constraints imposed by the χ^2 detector represent a convex set [91].

Next, we look at the MEWMA detector. We write the state of the MEWMA detector with zero as the initial condition as

$$x_D(k) = \sum_{i=\underline{k}}^{k-1} \beta(1-\beta)^{k-1-i} a(i) = T_\beta(k)^T a$$

such that the constraints to remain undetected become

$$y_D(k) = \frac{2-\beta}{\beta} ||T_\beta(k)^T a||_2^2 \leq J_D^M \quad \forall k \in \Gamma.$$

Here, we see again that $\frac{2-\beta}{\beta}||T_\beta(k)^T a||_2^2$ for each k is a convex function. Therefore, the union of all constraints represents a convex set.

Finally, let us investigate the constraint of the CUSUM detector. First, we prove that $y_D(0)$ is convex. Here, $y_D(0) = 0$ is given and constant. It is, therefore, simultaneously convex and concave in a. Now assume $y_D(k)$ is convex and let us prove that $y_D(k+1)$ is convex as well. We know that $||a(k)||_2^2$ is convex and, furthermore, $-\delta$ is convex because it is constant. Using [91], we obtain that the nonnegative weighted sum of convex functions is convex and taking the maximum of two convex functions results in a convex function as well. Hence, $y_D(k) + ||a(k)||_2^2 - \delta$ is convex and because of that $y_D(k+1) = \max\big(0, y_D(k) + ||a(k)||_2^2 - \delta\big)$ is also convex, which concludes the proof by induction that $y_D(k)$ represents convex constraints for all $k \in \Gamma$.

Hence, the constraints for the χ^2, CUSUM, and MEWMA detectors are convex sets. □

Now that we have shown that the constraint sets are convex, no matter which of the three detectors of interest we use, we can obtain the global solution to (6.3) by splitting Problem 6.1 into n_x subproblems.

Theorem 6.1. *The global solution of Problem 6.1 can be found by solving n_x convex optimization problems,*

$$\max_{i\in\{1,\cdots,n_x\}} \max_{a} t_{xa,i}^T a \quad \text{s.t.} \quad d\big(x_D(k), a(k)\big) \leq J_D \quad \forall\, k \in \Gamma,$$

where $t_{xa,i}^T$ is the ith row of T_{xa}.

Proof. Recall that the infinity norm can be written as

$$\max_{a} ||T_{xa}a||_\infty = \max_{a} \max_{i\in\{1,\cdots,n_x\}} |t_{xa,i}^T a|.$$

Hence, we can solve (6.3) by solving n_x separate problems and the maximum solution of the n_x problems is the solution of (6.3). Since both the χ^2 and CUSUM detectors use $||a(k)||_2^2$ to determine $y_D(k)$, $-a$ is a feasible solution if a is feasible. Similarly, since the MEWMA detector with zero as an initial condition has the output $y_D(k) = \frac{2-\beta}{\beta}||T_\beta(k)^T a||_2^2$ (see Proposition 6.1), $-a$ is also a feasible solution if a is feasible. Now we can show that the absolute value in the objective of each of the n_x problems can be removed. Assume we found two feasible solutions for one of the subproblems, $a_{\max}^*$ and $a_{\min}^*$, which lead to the maximum and minimum value of $t_{xa,i}^T a$ under the constraints, $x_{\max}^*$ and $x_{\min}^*$, respectively. We assume that $x_{\min}^* < 0 < x_{\max}^*$ and $x_{\max}^* < |x_{\min}^*|$. Since $-a_{\min}$ is also a feasible solution, we are able to define $a^* = -a_{\min}^*$ as a feasible solution, which leads to a higher value than $x_{\max}^*$. Hence, we do not need to consider the absolute value in each of the n_x problems. Further, $t_{xa,i}^T a$ is simultaneously concave and convex [91] and the constraints imposed by the χ^2, CUSUM, and MEWMA detectors are convex (Proposition 6.1). Hence, each of the n_x subproblems is convex, such that we obtain the global solution to Problem 6.1 by taking the maximum optimal value of the subproblems. □

Theorem 6.1 implies that we can find the global solution of the nonconvex optimization problem by solving n_x convex optimization problems. Therefore, we can find the worst-case impact of the attack (3.11) under the χ^2, CUSUM, and MEWMA detectors, which is necessary for the detector comparison.

Furthermore, if the χ^2 detector is used, the Problem 6.1 has an analytical solution.

Proposition 6.2. When a χ^2 detector is used, the optimization problem (6.3) becomes

$$\begin{aligned} \max_{a} ||x_a(\bar{k})||_\infty = \max_{a} \max_{i\in\{1,\cdots,n\}} |t_{xa,i}^T a| \\ \text{s.t. } a(k)^T a(k) \leq J_D \ \forall k \in \Gamma, \end{aligned} \tag{6.4}$$

where $t_{xa,i}^T = [t_{xa,i\underline{k}}^T, \cdots, t_{xa,i\bar{k}}^T]$ is the ith row of T_{xa} with $\bar{k} - \underline{k} + 1$ partitions $t_{xa,ik} \in \mathbb{R}^{n_y}$. The solution to (6.4) is

$$a^*(k) = (-1)^j \frac{\sqrt{J_D}}{\sqrt{t_{xa,i^*k}^T t_{xa,i^*k}}} t_{xa,i^*k} \tag{6.5}$$

and the attack impact is $\mathcal{I}(\tau) = \sqrt{J_D} \sum_{k=\underline{k}}^{\bar{k}} \sqrt{t_{xa,i^*k}^T t_{xa,i^*k}}$, where

$$i^* \in \arg \max_{i \in \{1,\cdots,n_x\}} \sum_{k=\underline{k}}^{\bar{k}} \sqrt{t_{xa,ik}^T t_{xa,ik}}$$

and j is either 1 or 2 for all k.

Proof. The proof of Theorem 6.1 shows us that we can neglect the absolute value in the inner maximization problem when finding the optimal solution. Hence, we just look at $t_{xa,i}^T a$ and find the optimal solution a^* for the inner problem under the constraints. Note that $t_{xa,i}^T a = \sum_{k=\underline{k}}^{\bar{k}} t_{xa,ik}^T a(k)$ and that the constraints $a(k)^T a(k) \leq J_D$ at each time step are independent of the previous attack signals. Hence, to obtain the worst-case attack we can solve $\bar{k} - \underline{k} + 1$ quadratically constraint linear programs of the form

$$\max_{a(k)} t_{xa,ik}^T a(k) \quad \text{s.t.} \quad a(k)^T a(k) \leq J_D, \tag{6.6}$$

which have the solution $\hat{a}(k) = \frac{\sqrt{J_D}}{\sqrt{t_{xa,ik}^T t_{xa,ik}}} t_{xa,ik}$. Inserting these $\hat{a}(k)$ in the inner maximization problem and solving for the optimal i we obtain $a^*(k)$ and the optimal objective value stated above. Due to the absolute value of the inner optimization problem $-a = -[a^*(\underline{k})^T, \cdots, a^*(\bar{k})^T]^T$ is also an optimal solution to the problem. □

Remark 6.1. Proposition 6.2 shows us that the attack impact for the objective $||x_a(\bar{k})||_\infty$ is $\sqrt{J_D}$ scaled by a plant specific constant. Therefore, the impact under a χ^2 detector will have a similar behavior for different plants in the metric of [25].

Before we move on to the comparison of the detectors, we present a reformulation of the CUSUM detector.

Reformulation of the CUSUM detector

We now introduce an equivalent reformulation of the CUSUM detector that does not use the non-smooth max operator, which leads to a better numerical implementation.

Proposition 6.3. For a given attacker's objective $f(a)$, the two optimization problems

$$\max_{a} f(a) \text{ s.t. } \begin{cases} y_D(k+1) = \max\left(0, y_D(k) + ||a(k)||_2^2 - \delta\right) \leq J_D^C \\ y_D(\underline{k}) = 0 \end{cases}, \tag{6.7}$$

and

$$\max_{a, \{\tilde{y}_D(k)\}_{k=\underline{k}}^{\bar{k}}} f(a) \text{ s.t. } \begin{cases} \tilde{y}_D(k+1) \geq 0 \\ \tilde{y}_D(k+1) \geq \tilde{y}_D(k) + ||a(k)||_2^2 - \delta \\ \tilde{y}_D(k+1) \leq J_D^C \\ \tilde{y}_D(\underline{k}) = 0 \end{cases} \tag{6.8}$$

for $k \in [\underline{k},\ \bar{k}]$ are equivalent in the sense that their optimal solutions a^*, if they exist, coincide.

Proof. First of all we can see that $y_D(k) \leq \tilde{y}_D(k),\ \forall k$, if a is fixed and feasible for both (6.7) and (6.8). First assume we obtained an optimal solution a^*_{CUSUM} for (6.7). This solution also fulfills the constraints of (6.8), since $\tilde{y}_D(k) = y_D(k) \leq J_D^C$, which makes a^*_{CUSUM} a feasible solution for (6.8). But by solving (6.8) directly we might find a solution a_r^* such that,

$$f(a_r^*) \geq f(a^*_{\text{CUSUM}}). \tag{6.9}$$

Assume now we found an optimal solution $(a_r^*, \{\tilde{y}_D^*(k)\}_{k=\underline{k}}^{\bar{k}})$ for (6.8) and we use $y_D(k) = \tilde{y}_D^*(k)$ in (6.7), where we pick a feasible sequence $\{\tilde{y}_D^*(k)\}_{k=\underline{k}}^{\bar{k}}$ for (6.7) by using the lower bounds $\tilde{y}_D^*(k+1) = \max\left(0, \tilde{y}_D^*(k) + ||a(k)||_2^2 - \delta\right)$, which does not change the value of the objective function of (6.8). Then $\left(a_r^*, \{\tilde{y}_D^*(k)\}_{k=\underline{k}}^{\bar{k}}\right)$ fulfills the constraints of (6.7) and is, therefore, a feasible solution for (6.7). But again we might find a solution by solving (6.7) directly such that,

$$f(a^*_{\text{CUSUM}}) \geq f(a_r^*). \tag{6.10}$$

Hence, the inequalities (6.9) and (6.10) imply $f(a^*_{\text{CUSUM}}) = f(a_r^*)$, which makes the problems equivalent and the reformulation valid. □

6.1.2 Detector comparison for two different processes

In the previous section, we showed that it is possible to find a solution to Problem 6.1. In this section, we solve Problem 6.1 for different mean times between false alarms τ to determine the metric proposed in [25] to compare the χ^2 detector, the CUSUM detector, and the MEWMA detector. Without loss of generality, we assume that $\underline{k} = 0$ and $\bar{k} = N$. Here, the attack will last for $N = 100$ time steps. Since the

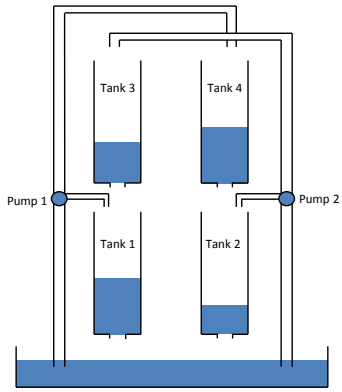

Figure 6.1: The quadruple tank process as proposed in [92]

impact depends on the process, we will determine the metric for the quadruple tank process and the three tank process, which was already used in Chapter 4.

The quadruple tank process is shown in Figure 6.1. After linearizing around the steady state reached for constant pump input voltages of 6 V, we obtain

$$\dot{x}(t) = \begin{bmatrix} -0.0723 & 0 & 0.1902 & 0 \\ 0 & -0.0633 & 0 & 0.1847 \\ 0 & 0 & -0.1902 & 0 \\ 0 & 0 & 0 & -0.1847 \end{bmatrix} x(t) + \begin{bmatrix} 0.1740 & 0 \\ 0 & 0.1506 \\ 0 & 0.0904 \\ 0.1044 & 0 \end{bmatrix} u(t) + w(t),$$

$$y(t) = \begin{bmatrix} 0.2000 & 0 & 0 & 0 \\ 0 & 0.2000 & 0 & 0 \end{bmatrix} + v(t).$$

We discretize the system with a sampling period $T_s = 0.5\,\text{s}$ and use a linear-quadratic-Gaussian design to obtain K, L, and Σ_r. Further, we assume that $w(k) \sim \mathcal{N}(0, 0.1I_4)$ and $v(k) \sim \mathcal{N}(0, 0.01I_2)$. The system has $n_y = 2$ sensors, which measure the water level in the lower tanks.

Here, we want to compare the χ^2 detector with the CUSUM and MEWMA detectors. Due to the forgetting factors, δ and β, different configurations for the CUSUM and MEWMA detector are possible. Therefore, we investigate the MEWMA detector for three forgetting factors, $\beta \in \{0.2, 0.4, 0.8\}$, and the CUSUM detector for $\delta = 2n_y = 4$. Solving Problem 6.1 for different τ and different detectors leads to the metric presented in Figure 6.2. We see that the attack impact is highest for the χ^2 detector. Hence, the internal states of the CUSUM and MEWMA detectors help mitigate the attack impact, even if the attacker knows the internal state. Comparing the CUSUM and MEWMA detectors, we see that the CUSUM detector with $\delta = 4$ has a lower attack impact than the MEWMA detector with $\beta = 0.8$ for most τ. However, if $\beta = 0.2$ or $\beta = 0.4$ the attack impact with the MEWMA detector is

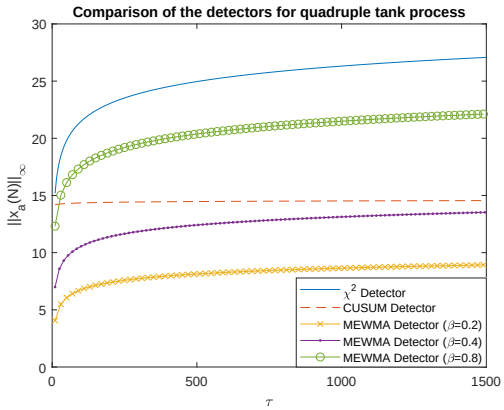

Figure 6.2: Impact of the attack (3.11) for a quadruple tank process over the mean time between false alarms τ.

lower for all investigated τ. Hence, the performance of the MEWMA detector can be better than the performance of the CUSUM detector. Note though that the impact for the CUSUM detector increases only slightly over the investigated τ, while the impact for the MEWMA detector increases more. Therefore, it looks like the impact under the MEWMA detector will be higher than for the CUSUM detector if we look at larger τ. Especially for $\beta = 0.4$, we observe that the impact is very close to the impact of the CUSUM detector for τ close to 1500 time steps in Figure 6.2 and will probably pass it for larger τ. Furthermore, we used $\delta = 4$ and the only condition for δ is that δ needs to be greater than n_y. Hence, we could use a smaller δ to mitigate the impact more, but there is a limit to how much we could mitigate the impact.

If we use the three tank process, we obtain the metric in Figure 6.3. We observe that Figure 6.3 is very similar to Figure 6.2. The impact is lower but the conclusion we can draw from Figure 6.3 is the same conclusion we can draw from Figure 6.2.

6.2 A New Metric for Detector Comparison

The metric proposed in [25] needs to be calculated for each plant separately and the operator needs to know the attacker's objective $f(a)$. In Chapter 5, we only showed how an attacker could get access to the detector state when the detector has linear dynamics. Therefore, we will show now that there exists time-invariant set $\mathcal{B}$ for each detector such that if $a(k) \in \mathcal{B}$ the attack remains undetected regardless of the detector state $x_D(k)$. This means that an attacker could still launch the attack (3.11), but without the knowledge of $x_D(k)$, the attack will not be as effective. We use the time-invariant set $\mathcal{B}$ to define a new metric for detector comparison, which is *independent* of the attacker's objective. Further, this new metric only depends on

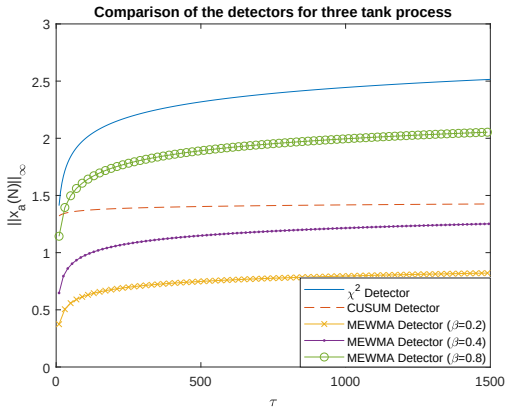

Figure 6.3: Impact of the attack (3.11) for a three tank process over the mean time between false alarms τ.

the number of sensors n_y and the detector used.

6.2.1 Time-invariant Sets for Stealthy Attacks

Let us first show that such a set $\mathcal{B}$ exists for all detectors that fulfills the conditions in Assumption 3.7.

Theorem 6.2. *There always exists a non-empty set*

$$\mathcal{B} = \{a(k) \in \mathbb{R}^{n_y} \mid a(k)^T a(k) \leq J\}$$

with $J > 0$ such that if $a(k) \in \mathcal{B}$ then $y_D(k+1) = d\big(x_D(k), a(k)\big) \leq J_D$ for all $x_D(k) \in \mathbb{X}$ if the detector fulfills the conditions in Assumption 3.7, where

$$\mathbb{X} = \{x \in \mathbb{R}^{n_D} : \ x = \theta(\tilde{x}, \tilde{r}), \text{where } \tilde{x} \text{ and } \tilde{r} \text{ are such that } d(\tilde{x}, \tilde{r}) \leq J_D\}. \quad (6.11)$$

Note that, $\mathbb{X}$ is the set of all possible detector states at time k when no alarm was triggered at k, i.e. $y_D(k) \leq J_D$. In case an alarm is triggered at k, the detector state is set to zero, i.e. $x_D(k) = 0$, and $0 \in \mathbb{X}$. Therefore, we only need to consider the possible detector states $x_D(k)$, when no alarm was triggered.

Proof. For the sake of readability, we will omit the time argument in this proof. We begin by showing that $\mathbb{X}$ represents a compact set. To prove the compactness of $\mathbb{X}$, let us first show

$$\{(x_D, a) \in \mathbb{R}^{n_D} \times \mathbb{R}^{n_y} : d(x_D, a) \leq J_D\} \quad (6.12)$$

is a compact set. Since $d\big(x_D, a\big)$ is assumed to be continuous and coercive in x_D and a (see Assumption 3.7), (6.12) is a compact set for all $J_D \geq 0$ (see Proof of

Corollary 2.5 in [93]). The set $\mathbb{X}$ is represented by a continuous map of all elements in the set (6.12) and is, therefore, compact, because a continuous map of a compact set is a compact set.

Now we choose

$$\epsilon = \min_{x_D \in \mathbb{X}} J_D - d(x_D, 0),$$

which, according to the extreme value theorem, exists, since $\mathbb{X}$ is compact and $d(x_D, 0)$ is continuous in x_D. Further, we know that $\epsilon > 0$, because $d(x_D, 0) < J_D$ for all $x \in \mathbb{X}$ (see second condition in Assumption 3.7).

Using the continuity of $d(x_D, a)$, we know that for the given $\epsilon > 0$ there exists a $\sqrt{J} > 0$ such that if $||a||_2 < \sqrt{J}$ then

$$\big|d(x_D, a) - d(x_D, 0)\big| < \epsilon.$$

This implies that $d(x_D, a) < d(x_D, 0) + \epsilon \leq J_D$ for all $x_D \in \mathbb{X}$ if $||a||_2 < \sqrt{J}$ due to the ϵ chosen.

This shows us that there exists an open non-empty ball $\tilde{\mathcal{B}}$ with radius $\sqrt{J} > 0$ in the set of all undetectable attack vectors a,

$$\mathbb{Z} = \big\{a \in \mathbb{R}^{n_y} : d(x_D, a) \in [0, J_D] \ \forall x \in \mathbb{X}\big\}.$$

Since $d(x_D, a)$ is continuous, $\mathbb{Z}$ is a closed set. Therefore, the closure of $\tilde{\mathcal{B}}$, which we denote as

$$\mathcal{B} = \big\{a \in \mathbb{R}^{n_y} \mid a^T a \leq J\big\},$$

is contained in $\mathbb{Z}$ as well. This concludes the proof. □

Note that, J depends only on the number of sensors n_y, the mean time between false alarms τ, since J_D is a function of τ, and the detector dynamics. Thus, J does not depend on the actual plant dynamics.

Since the χ^2, CUSUM, and MEWMA detectors fulfill the conditions in Assumption 3.7, we will show now how the set $\mathcal{B}$ looks like for these three detectors.

Attack Set for the χ^2 Detector

Since the χ^2 detector is a stateless detector, the attacker does not need to take the detector state $x_D(k)$ into account. We see that under attack $y_D(k+1) = a(k)^T a(k)$ is solely determined by $a(k)$ such that $a(k)^T a(k) > J_D^{\chi^2}$ would immediately cause an alarm. Hence, it follows that

$$\mathcal{B}^{\chi^2} = \big\{a(k) \in \mathbb{R}^{n_y} \mid a(k)^T a(k) \leq J_D^{\chi^2}\big\}, \tag{6.13}$$

which is already presented in [19] for a constant $a(k)$.

Attack Set for the MEWMA Detector

Recall that the MEWMA detector filters $r(k)$ and then looks at the size of the filtered $r(k)$ similar to the χ^2 detector. Under attack, the MEWMA detector dynamics are

$$\begin{aligned} x_D(k+1) &= (1-\beta)x_D(k) + \beta a(k) \\ y_D(k+1) &= ||x_D(k+1)||_2. \end{aligned}$$

Note that we rewrote the MEWMA detector a bit compared to Section 3.2.1, such that no alarm is triggered if $y_D(k+1) \leq \sqrt{\frac{\beta}{2-\beta} J_D^M}$.

Proposition 6.4. The largest time-invariant attack set under a MEWMA detector for an attacker that wants to remain undetected for all $k \geq \underline{k}$ and has no access to $x_D(k)$ is given by

$$\mathcal{B}^M = \Big\{ a(k) \in \mathbb{R}^{n_y} \;\Big|\; a(k)^T a(k) \leq \tfrac{\beta}{2-\beta} J_D^M \Big\}. \tag{6.14}$$

Proof. For $k \geq \underline{k}$ we can write $x_D(k)$ as

$$x_D(k) = (1-\beta)^{k-\underline{k}} x_D(\underline{k}) + \beta \sum_{i=0}^{k-\underline{k}-1} (1-\beta)^{k-\underline{k}-1-i} a(i+\underline{k}).$$

Let $||a(k)||_2 \leq \sqrt{J^M}$ for all $k \geq \underline{k}$. By using the triangle inequality on $y_D(k)$, we obtain

$$y_D(k) = ||x_D(k)||_2 \leq \big(1-(1-\beta)^{k-\underline{k}}\big)\sqrt{J^M} + (1-\beta)^{k-\underline{k}} y_D(\underline{k}) \leq \sqrt{\tfrac{\beta}{2-\beta} J_D^M},$$

where we used that $y_D(\underline{k}) = ||x_D(\underline{k})||_2$ and

$$\beta \sum_{i=\underline{k}}^{k-\underline{k}-1} (1-\beta)^{k-\underline{k}-1-i} = 1-(1-\beta)^{k-\underline{k}}.$$

To remain undetected independent of $y_D(\underline{k})$ one has to guarantee that

$$\begin{aligned} \sqrt{J^M} &\leq \min_{y_D(\underline{k}) \in [0, \sqrt{\frac{\beta}{2-\beta} J_D^M}]} \frac{\sqrt{\frac{\beta}{2-\beta} J_D^M} - (1-\beta)^{k-\underline{k}} y_D(\underline{k})}{1-(1-\beta)^{k-\underline{k}}} \\ &= \sqrt{\tfrac{\beta}{2-\beta} J_D^M} \end{aligned}$$

for all $k \geq \underline{k}$. Hence, $\mathcal{B}^M$ is the largest possible time-invariant set for a stealthy attack on all sensors without knowledge about the detector's state. □

Attack Set for the CUSUM Detector

As shown in Section 3.2.1, the CUSUM detector sums up the squared Euclidean norm of residuals with a forgetting factor. Further, the output equals the internal state of the CUSUM detectors such that the dynamics under attack are

$$y_D(k+1) = \max\big(y_D(k) + a(k)^T a(k) - \delta, 0\big).$$

Proposition 6.5. The largest time-invariant stealthy attack set under a CUSUM detector for an attacker, which has no access to $y_D(k)$ is given by

$$\mathcal{B}^C = \big\{a(k) \in \mathbb{R}^{n_y} \mid a(k)^T a(k) \leq \delta\big\}. \tag{6.15}$$

Proof. Assume the attack vector is limited by $a(k)^T a(k) \leq \delta + \epsilon$, where $\epsilon > 0$ is arbitrarily small. If $a(k)^T a(k) = \delta + \epsilon$ and $y_D(\underline{k}) \geq 0$, we get

$$y_D(\underline{k}+1) = \max\big(y_D(\underline{k}) + \epsilon, 0\big) = y_D(\underline{k}) + \epsilon \leq J_D^C$$

for the attack to remain undetected in its first step. Assume $y_D(\underline{k}) = J_D^C$ then this attack will trigger an alarm. If $y_D(\underline{k}) < J_D^C$ and $a(k)^T a(k) = \delta + \epsilon$ for all $k \geq \underline{k}$, an alarm is raised after $l = \left\lceil \frac{J_D^C - y_D(\underline{k})}{\epsilon} \right\rceil$ time steps. Hence, $\mathcal{B}^C$ defines the largest possible attack set for $a(k)$ that guarantees that the attack remains undetected for all $k \geq \underline{k}$ no matter the value of $y_D(\underline{k})$. □

6.2.2 A Detector Metric for Sensor Attacks

In the previous section, we showed that even if the detector has an internal state, the attacker is able to launch a stealthy attack. Hence, just having a detector with an internal state does not mean that a stealthy attack is impossible, but it might still make the system more secure, which is what we want to investigate here.

In case the attacker does not know the internal state $x_D(k)$ of the detector, we can redefine the attack impact as

Definition 6.1. The worst-case impact of the stealthy attack (3.11) without knowledge of the detector state on the closed-loop system (3.12) with zero initial conditions, no noise ($w(k) = 0$), and equipped with an anomaly detector (3.3) defined as

$$\bar{\mathcal{I}} := \sup_a f(a) \text{ s.t. } ||a(k)||_2^2 \leq J \ \forall k \geq \underline{k}, \tag{6.16}$$

where $a = \{a(k)\}_{k \geq \underline{k}}$ represents the attack trajectory and $f(a)$ characterizes the attacker's impact according to a certain objective.

From the perspective of the defender, we are not able to know the attacker's objective $f(a)$, when we design our system and choose the anomaly detector. In the following, we show that for attackers without knowledge of $x_D(k)$ we are able to compare the ability of each detector to mitigate the worst-case attack impact *independent* of the details of the attack objective $f(a)$.

Theorem 6.3. *The worst-case impact of the stealthy attack $\bar{\mathcal{I}}$ is non-decreasing in J, i.e. $\bar{\mathcal{I}}(J_1) \geq \bar{\mathcal{I}}(J_2)$ if $J_1 \geq J_2$.*

Proof. We see that the domain of $a(k)$ grows with J. Therefore, $||a(k)||_2^2 \leq J_1$ includes $||a(k)||_2^2 \leq J_2$ if $J_1 > J_2$ and the attack impact for J_1 is no smaller than the one obtained with J_2. □

Now we use Theorem 6.3 to compare the detectors' performance under attack, i.e. which detector mitigates the attack impact the most for a certain mean time between false alarms.

Remember for detector comparison in [25], we need to plot the worst-case impact over the mean time between false alarms τ. For the metric in [25] we need to solve the optimization problem in Definition 6.1 for each J to obtain the attack impact. The impact depends on the plant investigated as Figures 6.2 and 6.3 showed. Therefore, the metric in [25] is plant specific and needs to be recalculated for each plant.

Using Theorem 6.3, we see that the attack impact $\bar{\mathcal{I}}(J)$ is non-decreasing in J. Thus, instead of $\bar{\mathcal{I}}(J)$ we compare J over τ for each detector. Recall that J for each detector is a function of n_y, τ, and and the detector dynamics. This means that the metric for detector comparison under full sensor attacks that we propose depends only on the number of sensors in the plant and not on the actual plant dynamics. Therefore, the metric for a certain n_y applies to *all* plants with n_y sensors.

In the following, we compare only the χ^2, CUSUM, and MEWMA detectors. However, our proposed metric applies to all detectors that fulfill Assumption 3.7. For these three detectors we know that

$$J = \begin{cases} J_D^{\chi^2} & \text{for the } \chi^2 \text{ detector,} \\ \delta & \text{for the CUSUM detector,} \\ \frac{\beta}{2-\beta} J_D^{\mathrm{M}} & \text{for the MEWMA detector.} \end{cases}$$

Figure 6.4 shows how the metric looks like for a system with $n_y = 2$ sensors, for $\delta = 2n_y = 4$, $\beta \in \{0.2, 0.4, 0.8\}$, and $\tau \in [10, 1500]$. We can directly see that J for the CUSUM and MEWMA detectors is smaller than J for the χ^2 detector. Therefore, we conclude that the investigated stateful detectors mitigate the impact of the worst-case attacks without knowledge of $x_D(k)$ more than the χ^2 detector for *any* attack objective $f(a)$.

Recall that the only constraint for δ is $\delta > n_y$ to obtain a stochastically stable CUSUM detector. If we, for example, had chosen $\delta = 15$, the CUSUM detector might lead to a larger attack impact than the χ^2 detector. Therefore, we can use this metric to determine an upper bound for δ such that the CUSUM detector is stochastically stable and leads to a lower worst-case attack impact than the χ^2 detector. Furthermore, note that δ is constant over all τ such that the worst-case impact under a CUSUM detector does not depend on τ in contrast to the other detectors investigated. This makes the CUSUM detector attractive for an operator, since an operator desires a large time between false alarms and a low attack impact.

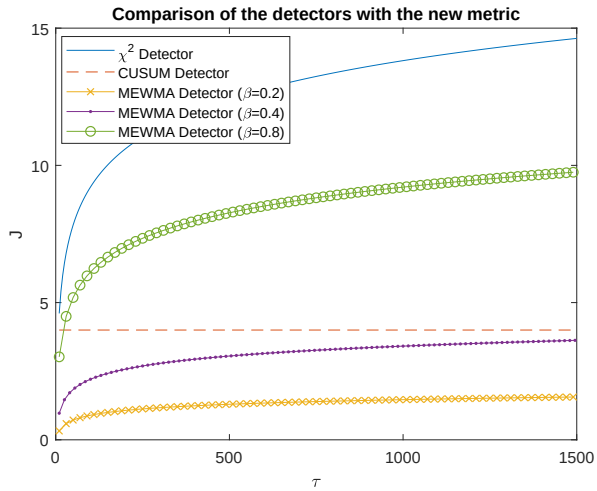

Figure 6.4: Our proposed metric for detector comparison, which plots J over τ to assess the attack impact at a certain τ

Furthermore, we observe that the MEWMA detector has an even lower J than the CUSUM and χ^2 detectors for $\beta \in \{0.2, 0.4\}$ and $\tau \leq 1500$ time steps, while J for $\beta = 0.8$ is bigger than δ under the CUSUM detector for $\tau > 20$ time steps. This means that the MEWMA detector might lead to a smaller attack impact than the CUSUM detector for $\beta \in \{0.2, 0.4\}$. Thus, with the right configuration the MEWMA detector is a suitable alternative to the CUSUM detector for these attacks.

Remark 6.2. The approximated thresholds for the MEWMA detector suggest that $\frac{\beta}{2-\beta} J_D^M \leq J_D^{\chi^2} \; \forall \beta \in (0, 1]$. Since the worst-case impact is non-decreasing in J according to Theorem 6.3 this would mean that the impact of the attacks without knowledge of $x_D(k)$ under the MEWMA detector is never larger than the impact under the χ^2 detector for the investigated attack strategy. However, we were not able to prove that the above inequality holds for all $\beta \in (0, 1]$.

If we compare Figure 6.2 and Figure 6.3 with Figure 6.4, we observe that our proposed metric gives us a similar result to the metric proposed in [25]. In Figures 6.2 and 6.3 the attack impact for the CUSUM detector increases while it is constant according to our proposed metric. This is due to the fact that for our proposed metric the attacker does not use any knowledge about the internal state of the detector and δ does not depend on τ. However, the impact for the CUSUM detector increases only slowly with τ. Hence, the constant J in our metric is a good approximation for how the impact under the CUSUM detector changes. Our metric also agrees with the metric of [25] that a stateless detector performs worse than a detector with an internal state for the investigated forgetting factors.

Remark 6.3. It is important to note that we have only calculated the metric proposed in [25] for the case where the attacker's objective is to maximize the infinity norm of $x_a(\bar{k})$. In this case, our proposed metric and the metric proposed in [25] give similar results on the performance of the detectors. It is, therefore, important to also investigate how the performance for the metric in [25] changes when another attack objective is used.

6.3 Summary

In this chapter, we compared the performance of the χ^2, CUSUM, and MEWMA detectors using the metric proposed in [25]. The comparison showed us that detectors with internal dynamics mitigate the attack impact more than static detectors. Further, the performance of the CUSUM and MEWMA detectors depends on the forgetting factor chosen and for certain forgetting factors the MEWMA detector performs better than the CUSUM detector over the range of investigated τ. Moreover, we showed that there exists a time-invariant set, such that if the attack vector is in this set, the attack remains undetected regardless of the internal state of the detector. This time-invariant set was then used to propose a new metric for detector comparison that does not depend on the attacker's objective, but only on the number of sensor in the plant. This new metric yields very similar result in the detector comparison as the metric proposed in [25]. However, once we compared the performance of different detectors and picked the detector that performs best, we still need to decide which threshold we want to use. This corresponds to choosing a mean time between false alarms that is optimal with respect to a certain objective. A way to optimally choose the mean time between false alarms is proposed in the next chapter.

Chapter 7

A Game-Theoretic Approach to Detector Tuning

In the previous chapter, we used the metric proposed in [25] to compare the performance of the χ^2, CUSUM, and MEWMA detectors. We also proposed a new metric, which does not depend on the attacker's objective and the plant dynamics, to compare the performance of detectors. However, the metric of Urbina *et al.* [25] and our proposed metric presented in Chapter 6 only help us to decide which detector performs best. These metrics do not help us to choose J_D in an optimal way when an attacker is present. Therefore, this chapter introduces a game-theoretic framework to pick the optimal J_D for a detector. This game-theoretic framework is an adaption of the framework of Ghafouri *et al.* [26] to our scenario of a stealthy sensor attack. The basic idea is to minimize the sum of the cost induced by the false alarms and by the attack. In [26], the attacks are detectable with a delay depending on the detector tuning and it is not specified how the impact of the attack can be obtained in an analytical manner. Therefore, we analyze the Stackelberg game formulation of [26] in a more control-theoretic context. Here, we use it to find a fixed detector threshold in the presence of the stealthy sensor attack presented in Chapter 3. We show that the Stackelberg game used in our work always has a solution and present sufficient conditions for the uniqueness of the solution. Further, we show that the optimal action of the attacker in this game represents the detector metric presented in [25] and how the Stackelberg game complements the metric in [25]. This Stackelberg framework can be seen as a tool to treat worst-case attackers with knowledge about the plant.

Note that, for the results in this chapter to hold, we only need the detector functions to be continuous in their arguments, $d\big(x_D(k), a(k)\big)$ to be coercive in its arguments, $\theta(0,0) = 0$, and $d(0,0) = 0$. Hence, these results also hold for a less strict detector model than we assumed in Chapter 3.

Let us restate the definition for the worst-case attack impact of a time-limited attack.

Definition 7.1. The worst-case impact $\mathcal{I} : \mathbb{L} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ of the stealthy sensor

attack (3.11) on the closed-loop system (3.12) equipped with an anomaly detector (3.3) is defined as

$$\mathcal{I}(\tau) := \max_{a, x_D(\underline{k})} f(a) \quad \text{s.t.} \quad d\big(x_D(k), a(k)\big) \leq J_D \quad \forall\, k \in \Gamma,$$

where $a = \{a(k)\}_{k=\underline{k}}^{\bar{k}}$ is the attack trajectory and $f(a)$ is a continuous function that characterizes the attacker's objective.

Before we can move on to the solution of the problem, we need to familiarize ourselves with the concept of a Stackelberg game.

7.1 Stackelberg Games

A two-player Stackelberg game consists of a leader, who plays first, and a follower, who plays second and is aware of the action taken by the leader. Both players are assumed to be rational, which means that they always try to minimize/maximize their respective cost/payoff. Let us denote the action of the leader by $m_L \in M_L$, where M_L contains all possible actions of the leader. The action of the follower is denoted by $m_F \in M_F$, where M_F contains all possible actions of the follower.

The goal of the leader is to minimize its cost $c_L(m_L, m_F)$, while the follower wants to maximize its payoff $p_F(m_L, m_F)$. We assume that the leader knows M_F and $p_F(m_L, m_F)$, while the follower knows the leader's action m_L. What is good for the follower is not necessarily good for the leader, e.g. increasing the follower's payoff $p_F(m_L, m_F)$ might also increase the leader's cost $c_L(m_L, m_F)$. Hence, the leader has to pick its action so that it has the minimum cost for all actions possible of the follower. One common solution concept used for Stackelberg games is the subgame perfect Nash equilibrium (SPNE), see for example [94]. The SPNE is typically obtained via backwards induction. First, the optimal actions m_F^* of the follower to maximize its payoff depending on m_L are obtained. Then, given the dependency of m_F^* on m_L we can obtain the optimal action of the leader m_L^*. In case m_F^* is not unique for a given m_L and results in different costs for the leader, the leader wants to minimize the least upper bound of its cost [95]. Hence, the SPNE of the two-player Stackelberg game can be formulated as the following two-level optimization problem

$$\min_{m_L \in M_L} \sup_{m_F^*} c_L(m_L, m_F^*) \quad \text{s.t.} \quad m_F^* \in \arg\max_{m_F' \in M_F} p_F(m_L, m_F'). \tag{7.1}$$

A solution (m_L^*, m_F^*) to (7.1) is called a generalized Stackelberg strategy pair [95]. As stated in [95], (7.1) can also be interpreted as a follower who wants to maximize its payoff but also maximize the leader's cost if possible.

7.2 Finding the Optimal Tuning

In this section, we first investigate the existence of the worst-case impact $\mathcal{I}(\tau)$, then formulate the Stackelberg game to find the optimal detector threshold J_D and

present the attacker's and defender's optimal actions. Finally, we investigate when the solution to the Stackelberg game is unique.

7.2.1 Existence of Solutions to the Worst-case Impact

We begin by utilizing the extreme value theorem to show that the attack impact $\mathcal{I}(\tau)$ exists. Recall from Assumption 3.8 that $\mathbb{L}$ is the compact set of mean time between false alarms.

Theorem 7.1. *The worst-case attack impact $\mathcal{I}(\tau)$ is well-defined on $\mathbb{L}$ and is non-decreasing in τ.*

Proof. Let us first show that $\mathcal{I}(\tau)$ is well-defined on $\mathbb{L}$. The constraint set in Definition 7.1 is non-empty, because $a = 0$ and $x_D(\underline{k}) = 0$ fulfill the constraints independent of how J_D is chosen, since then $x_D(k) = 0$ for all $k \in \Gamma$ and $d(0, 0) = 0$ (see Assumption 3.7).

According to the extreme value theorem, an optimization problem over a continuous function has at least one solution if the constraints represent a compact set. To prove the compactness of the constraint set, recall that the first constraint in Definition 7.1,

$$\{(x_D(k), a(k)) \in \mathbb{R}^{n_D} \times \mathbb{R}^{n_y} : d(x_D(k), a(k)) \leq J_D\}, \tag{7.2}$$

is a compact set (see proof of Theorem 6.2).

The remaining constraints can be written as

$$\begin{aligned} d\big(\theta(x_D(\underline{k}), a(\underline{k})), a(\underline{k}+1)\big) &\leq J_D \\ d\Big(\theta\big(\theta(x_D(\underline{k}), a(\underline{k})), a(\underline{k}+1)\big), a(\underline{k}+2)\Big) &\leq J_D \\ &\vdots \end{aligned} \tag{7.3}$$

Since $\theta(x_D(k), a(k))$ is continuous in its arguments and the composition of continuous functions is continuous, we note that the functions on the left hand side of the inequalities in (7.3) are continuous in their arguments. Therefore, each of the constraints in (7.3) represents a closed set [96]. Finally, since the intersection of compact and closed sets is compact, the constraint set in Definition 7.1 represent a compact set. Hence, $\max_{a, x_D(\underline{k})} f(a)$ subject to the constraints exists for all $J_D = g(\tau) \geq 0$ and it is, therefore, well-defined on $\mathbb{L}$.

Now let us show that $\mathcal{I}(\tau)$ is non-decreasing in τ. Since $J_D = g(\tau)$ is non-decreasing (by Assumption 3.8) and the domain is growing with J_D, the attack impact $\mathcal{I}(\tau)$ is non-decreasing in τ. □

Theorem 7.1 shows that the worst-case impact might increase for our general class of detectors if we want to reduce the number of false alarms of our detector by increasing J_D. For the remainder of this chapter, we make the following assumption.

Assumption 7.1. The worst-case impact of the stealthy attack $\mathcal{I}(\tau)$ is continuous on $\mathbb{L}$.

7.2.2 Stackelberg Formulation

We present a Stackelberg game to find the optimal detector tuning, where the defender's action is to choose τ, while the attacker's action is to choose $\{a, x_D(\underline{k})\}$. Since the attacker knows $x_D(k)$, maximizing over $x_D(\underline{k})$ can be interpreted as the attacker waiting for the most opportune moment to launch its attack.

From a defender's perspective, every alarm raised by the detector has to be investigated to check if an attack is happening. Let $c_{\mathrm{FA}} \geq 0$ be the cost for investigating a false alarm and $c_{\mathrm{A}} > 0$ be the cost factor of an attack. Similar to [26], the cost c_L, a defender wants to minimize, is

$$c_L(\tau, \{a, x_D(\underline{k})\}) = c_{\mathrm{FA}} \frac{T}{T_s \tau} + c_A f(a), \tag{7.4}$$

where $T_s \in \mathbb{R}_{>0}$ is the sampling period, and $T \in \mathbb{R}_{>0}$ is the length of a time interval of interest for the defender. The first term represents the cost induced by all false alarms during the time interval T, since $\frac{T}{T_s}$ are the number of samples taken over T and τ is the average time between false alarm in samples. The second term represents the cost induced by the stealthy attack. From the attacker's perspective the goal is to have the highest impact possible, while remaining undetected. Hence, in the notation of Section 7.1 we have

$$p_F(\tau, \{a, x_D(\underline{k})\}) = f(a) \quad \text{s.t.} \quad d\big(x_D(k), a(k)\big) \leq J_D \quad \forall k \in \Gamma. \tag{7.5}$$

Now, we can formulate the problem of finding the optimal detector tuning, which minimizes the cost c_L as a Stackelberg game, where the defender is the leader and the attacker is the follower. This order is intuitive since the defender has to first set up the defenses before an attacker can penetrate them.

Problem 7.1. The optimal detector tuning under stealthy sensor data attacks according to the specified costs and payoffs above is characterized by the Stackelberg game

$$\min_{\tau \in \mathbb{L}} \; c_{\mathrm{FA}} \frac{T}{T_s \tau} + c_A f(a) \quad \text{s.t.} \quad \{a, x_D(\underline{k})\} \in \arg \max_{\{a', x'_D(\underline{k})\} \in \mathbb{A}} f(a'),$$

where

$$\mathbb{A} = \big\{ \big(a, x_D(\underline{k})\big) \in \mathbb{R}^{(\bar{k}-\underline{k}+1)n_y} \times \mathbb{R}^{n_D} : \quad d\big(x_D(k), a(k)\big) \leq J_D \quad \forall k \in \Gamma \big\}.$$

Note that we do not need the supremum operator of (7.1) in Problem 7.1, because even if $\{a, x_D(\underline{k})\}$ is not unique for a given τ, each $\{a, x_D(\underline{k})\}$ will have the same influence on c_L because $f(a)$ is unique.

Now we will look at each player and present existence and uniqueness results for Problem 7.1. As it is common for finding the SPNE, we will first look at the follower's (attacker's) action and then at the leader's (defender's) action.

7.2.3 The Attacker's Action

Since the attacker wants to maximize the attack impact, its set of possible worst-case attacks is defined as

$$\mathbb{A}^* := \arg \max_{\{a', x'_D(\underline{k})\} \in \mathbb{A}} f(a'). \tag{7.6}$$

Corollary 7.1. *The set of stealthy worst-case attacks $\mathbb{A}^*$ is non-empty for all $J_D \geq 0$.*

Proof. The proof of Theorem 7.1 showed us that there always exists at least one solution to the optimization problem stated in Definition 7.1. Therefore it follows that $\mathbb{A}^*$ is non-empty. □

By Definition 7.1 the value $\mathcal{I}(\tau) = f(a)$ $\forall \{a, x_D(\underline{k})\} \in \mathbb{A}^*$ is unique even if we have more than one attack in $\mathbb{A}^*$.

Note that if we plot $\mathcal{I}(\tau)$ over τ, we get the metric presented in [25]. Hence, there is a close relation between the attacker's action in our Stackelberg game and the metric proposed in [25].

7.2.4 The Defender's Action

Given that the attacker will play $\{a, x_D(\underline{k})\} \in \mathbb{A}^*$ the minimization problem becomes

$$\min_{\tau \in \mathbb{L}} \; c_{FA} \frac{T}{T_s \tau} + c_A \mathcal{I}(\tau) \tag{7.7}$$

and we are able to show the following about the solutions of (7.7).

Theorem 7.2. *There exists at least one solution to* (7.7).

Proof. Since $c_{FA} \frac{T}{T_s \tau} + c_A \mathcal{I}(\tau)$ is continuous on $\mathbb{L}$ and $\mathbb{L}$ is a compact set, we always have at least one minimum in $\mathbb{L}$ due to the extreme value theorem. □

The result above shows us that there exists at least one solution to (7.7). We now give two sufficient conditions for uniqueness of the solution.

Proposition 7.1. Let $\mathcal{I}(\tau)$ be twice continuously differentiable in τ, and let $\mathcal{I}'(\tau)$ and $\mathcal{I}''(\tau)$ denote the first and second derivative of $\mathcal{I}(\tau)$ with respect to τ, respectively, then the solution to (7.7) is unique if

$$\mathcal{I}''(\tau) > \frac{-2}{\tau} \mathcal{I}'(\tau) \quad \forall \tau \in \mathbb{L}. \tag{7.8}$$

Proof. Let $z = \frac{1}{\tau}$ and $h(z) = \mathcal{I}(\frac{1}{z})$, where $z \in \mathbb{K} := \{\alpha \in [0, 1] : \alpha = \frac{1}{\tau} \, \forall \tau \in \mathbb{L}\}$. With that we can reformulate (7.7) to

$$\min_{z \in \mathbb{K}} \; c_{FA} \frac{T}{T_s} z + c_A h(z). \tag{7.9}$$

This has a unique solution if $c_{FA}\frac{T}{T_s}z + c_A h(z)$ is strictly convex in z. Here, $c_{FA}\frac{T}{T_s}z + c_A h(z)$ is strictly convex if $c_A h(z)$ is strictly convex. Since $c_A > 0$, the strict convexity of $c_A h(z)$ is guaranteed if $h''(z) > 0 \, \forall z \in \mathbb{K}$. Now with $h(z) = \mathcal{I}(\frac{1}{z})$ we get

$$\begin{aligned} h''(z) = \mathcal{I}''\Big(\frac{1}{z}\Big)\frac{1}{z^4} + 2\mathcal{I}'\Big(\frac{1}{z}\Big)\frac{1}{z^3} > 0 \quad \forall z \in \mathbb{K} \\ \Leftrightarrow \mathcal{I}''(\tau) > \frac{-2}{\tau}\mathcal{I}'(\tau) \quad \forall \tau \in \mathbb{L}, \end{aligned}$$

which concludes the proof. □

Note that the condition in Proposition 7.1 is independent of c_{FA}, T, T_s, and c_A. Therefore, it does not depend on the exact parametrization of the cost function (7.7). Let us now present another sufficient condition for the uniqueness of the solution of (7.7), which does not depend on $\mathcal{I}'(\tau)$ but on c_{FA}, T, T_s, and c_A.

Proposition 7.2. Let $\mathcal{I}(\tau)$ be twice continuously differentiable in τ, and let $\mathcal{I}''(\tau)$ denote the second derivative of $\mathcal{I}(\tau)$ with respect to τ, then the solution to (7.7) is unique if

$$\mathcal{I}''(\tau) > -2\frac{c_{FA}T}{c_A T_S}\frac{1}{\tau^3} \quad \forall \tau \in \mathbb{L}. \tag{7.10}$$

Proof. The solution of (7.7) is unique if $c_{FA}\frac{T}{T_s\tau} + c_A\mathcal{I}(\tau)$ is strictly convex, which is the case if its second derivative is greater than zero for all $\tau \in \mathbb{L}$. This leads to

$$2c_{FA}\frac{T}{T_s\tau^3} + c_A\mathcal{I}''(\tau) > 0, \tag{7.11}$$

which gives us the condition stated above. □

Note that both uniqueness conditions of Propositions 7.1 and 7.2 include all strictly convex $\mathcal{I}(\tau)$, but they also hold for some non-convex $\mathcal{I}(\tau)$, because $\mathcal{I}''(\tau) < 0$ is also possible.

Since we have stated two sufficient conditions for the uniqueness of the solution of (7.7), let us now investigate when the condition in Proposition 7.1 is more strict than the condition in Proposition 7.2.

Corollary 7.2. *The condition stated in Proposition 7.1 is less strict than the condition in Proposition 7.2 if*

$$\mathcal{I}'(\tau) > \frac{c_{FA}T}{c_A T_s}\frac{1}{\tau^2} \quad \forall \tau \in \mathbb{L}$$

and otherwise if

$$\mathcal{I}'(\tau) < \frac{c_{FA}T}{c_A T_s}\frac{1}{\tau^2} \quad \forall \tau \in \mathbb{L}.$$

Proof. Comparing the bounds on $\mathcal{I}''(\tau)$ in Proposition 7.1 and Proposition 7.2 readily leads to these conditions. □

This shows that Proposition 7.1 is less strict for impacts $\mathcal{I}(\tau)$ with a large slope, while Proposition 7.2 is less strict for impacts that increase slower in τ.

Note that if the conditions stated in Corollary 7.2 are not fulfilled for all $\tau \in \mathbb{L}$, we cannot make a general statement if the condition in Proposition 7.1 is stricter than that in Proposition 7.2 or not.

Recall that the metric in [25] compares the worst-case stealthy attack impacts under different detectors over the mean time between false alarms. If one detector results in a lower attack impact than another detector for some investigated same mean times between false alarms, we consider this detector better for these mean times between false alarms. If we were to compare several detectors and picked the best detector according to this metric, then we already have the worst-case stealthy attack impacts for different τ, i.e. we have $\mathcal{I}(\tau)$. This can immediately be used to solve (7.7) and obtain an optimal tuning for the chosen detector. Hence, Problem 7.1 complements the metric presented in [25].

7.3 Illustrative Example

Let us now present an illustrative example on how one can use the presented Stackelberg formulation to find an optimal tuning $J_D(\tau)$. For this example, we use the χ^2 detector, which, for the reader's convenience, is restated below,

$$y_D(k+1) = ||r(k)||_2^2 \leq J_D. \tag{7.12}$$

Here, $y_D(k)$ does not depend on $x_D(k)$, therefore in our Stackelberg game framework the attacker's action is represented only by a. We also see that (7.12) is coercive and continuous in $r(k)$ and, thus, fulfills Assumptions 3.7. Recall from Chapter 3 that

$$J_D = g(\tau) = 2P^{-1}\Big(\frac{n_y}{2}, 1 - \frac{1}{\tau}\Big), \tag{7.13}$$

where $P^{-1}(\cdot,\cdot)$ represents the inverse regularized lower incomplete gamma function, which is non-decreasing in τ and therefore fulfils Assumption 3.8.

Now that we have specified the detector, let us look at our attacked system in the second stage of the attack. Here, we consider an observer-based controller. Recall from Chapter 3 that we can express the closed-loop system as a superposition of two linear subsystems. One system is affected by the noise, while the other is affected by the attack. To investigate the attack impact we consider the attacked subsystem for all $k \geq \underline{k}$, which is describe by

$$\begin{aligned} x_a(k+1) &= (A - BK)x_a(k) + BKe_a(k) \\ e_a(k+1) &= Ae_a(k) - L\Sigma_r^{\frac{1}{2}}a(k) \end{aligned} \tag{7.14}$$

with $x_a(\underline{k}) = e_a(\underline{k}) = 0$.

The attack lasts N time steps. Without loss of generality let $\underline{k} = 0$ such that $\Gamma = [0, N-1]$ in this case. Let $a = [a(0)^T, \cdots, a(N-1)^T]^T$ and the attacker's target is $x_a(N)$. Since we consider a linear discrete-time system and $x_a(0) = 0$, we can obtain a matrix $T_{xa} \in \mathbb{R}^{n_x \times N n_y}$ from (7.14), such that $x_a(N) = T_{xa}a$. In the previous chapter, we used the infinity norm as the attacker's objective, but here we look at the squared infinity norm, i.e. $f(a) = ||T_{xa}a||_\infty^2$. Note that the worst-case attack strategy will be the same since the objective functions are equivalent under a monotone transformation. However, with the squared infinity norm the impact will depend directly on J_D, which leads to a more convenient notation. Recall that in the attack strategy $r(k) = a(k)\ \forall k \in \Gamma$. Hence, the constraints of the attacker to remain undetected by the χ^2 detector become

$$a(k)^T a(k) \leq J_D \quad \forall k \in \Gamma. \tag{7.15}$$

Note that, in this example, $f(a)$ is continuous and the constraints to remain undetected (7.15) represent a compact set. Hence, according to Theorem 7.1 the impact $\mathcal{I}(\tau)$ is well defined. By using Proposition 6.2 we obtain the worst-case attack impact as

$$\mathcal{I}(\tau) = J_D \sum_{k=0}^{N-1} \big(\sqrt{t_{xa,i^*k}^T t_{xa,i^*k}}\big)^2.$$

Here, the worst-case attack impact is a linear function of J_D. Hence, the solution of Problem 7.1 can be found as

$$\begin{aligned} \tau^* &= \arg\min_{\tau \in \mathbb{L}} c_{\mathrm{FA}} \frac{T}{T_s \tau} + c_J J_D \\ &= \arg\min_{\tau \in \mathbb{L}} c_{\mathrm{FA}} \frac{T}{T_s \tau} + c_J 2P^{-1}\Big(\frac{n_y}{2}, 1 - \frac{1}{\tau}\Big), \end{aligned} \tag{7.16}$$

where $c_J = c_{\mathrm{A}}\big(\sum_{k=0}^{N-1} \sqrt{t_{xa,i^*k}^T t_{xa,i^*k}}\big)^2$.

To give an illustration of the Stackelberg framework we look at the quadruple tank process equipped with a χ^2 detector. Recall that the quadruple tank process was introduced in Chapter 6. To obtain K, L, and Σ_r we use a linear-quadratic-Gaussian design, for which we linearize the system again around the steady state reached for constant input voltages of 6 V and we discretize the system with a sampling period $T_s = 0.5$ s. We consider a time horizon of a day, i.e. $T = 86400$ s. The cost for investigating a false alarm is $c_{FA} = 25$ and the cost of an attack is $c_A = 100$. Assuming that the attack lasts for $N = 1000$ time steps we obtain $c_J = 553.9848$.

Let $\mathbb{L} = [1, 10^6]$, then the Stackelberg framework gives us $\tau^* = 3899$ time steps. The upper plot in Figure 7.1 shows how the cost function behaves for different τ and we see that there exists a unique minimum. Let us compare τ^* with two

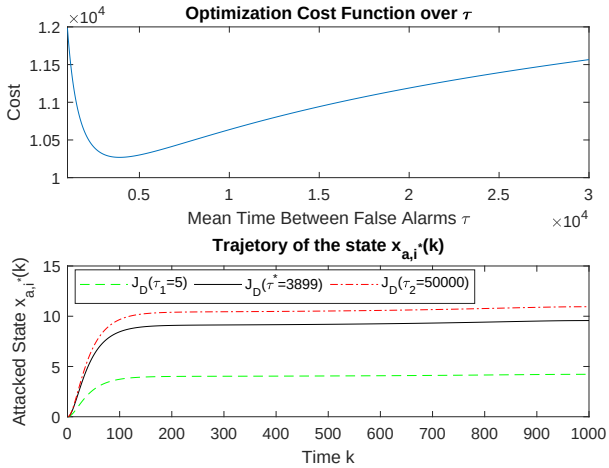

Figure 7.1: The upper plot shows the existence of a unique minimum of the cost function, while the lower plot depicts how the attack impact increases with τ.

extreme cases, $\tau_1 = 5$ and $\tau_2 = 50000$. The lower plot of Figure 7.1 illustrates the trajectory of the state, which the attacker can deteriorate the most. We see that the impact increases with τ. Here, $\tau_1 = 5$ has the lowest impact but then we have many false alarms, which is represented in the cost function. Therefore, we compare the analytical cost function and the average cost function over 10000 simulations. The analytical costs for τ_1, τ^*, and τ_2 are 865783, 10269, and 12074, respectively. Averaging the cost over 10000 simulations gives us 865793, 10269, and 12075 for τ_1, τ^*, and τ_2, respectively. We see that the average simulated cost for each investigated τ is similar to its analytically determined cost. Hence, this verifies that the proposed Stackelberg framework gives us the cost optimal average time between false alarms.

7.4 Summary

In this chapter, we proposed a Stackelberg game framework to tune an anomaly detector in the presence of sensor attacks. In this framework, the defender acts as a leader and chooses the detector threshold, while the attacker acts like a follower and execute its stealthy attack based on the detector threshold. The attacker tries to maximize its impact, while the defender chooses the threshold to minimize its cost. The cost is the sum of the cost for investigating false alarms and the cost the attack impact induces. We presented optimal strategies for both the attacker and defender and showed that a solution to the Stackelberg game exists. Further, two conditions for the uniqueness of the solution were given and we indicated a close connection between the Stackelberg game and a metric for detector comparison. An illustrative example showed how this Stackelberg framework is used to tune a χ^2 detector.

Chapter 8

Conclusions and Future Work

In this chapter, we will conclude the thesis and present several possible directions for future work.

8.1 Conclusions

In this thesis, we were concerned with the performance of anomaly detectors under stealthy sensor attacks. The mitigation of the attack impact was one of the main criteria for evaluating the performance of the detector.

Chapter 3 introduced our model of the sensor attack scenario. We used a linear plant and a linear controller to model the closed-loop system, which is equipped with an anomaly detector. The anomaly detector model we proposed is very general and could fit also possible nonlinear detector models. We fit three detectors, namely the χ^2, CUSUM, and MEWMA detectors, into this model and present tuning methods for them to achieve a certain mean time between false alarms. We then give an example how the metric proposed in [25] can be used to compare detectors. After presenting the system model, we discussed the attacker model including the assumption made on the attacker's resources and its goal. The goal is to remain undetected while maximizing a function, which corresponds to the attacker's objective. We introduced one possible worst-case attack strategy, which we used throughout the thesis. We pointed out that is not immediately possible to execute the worst-case attack when the attacker penetrates the network. The problem for the attacker is that it does not have knowledge of the controller's and detector's state at the beginning of the attack. This sets the stage for a feasibility analysis of this attack.

Chapter 4 started the feasibility analysis of the considered worst-case attack. Here, we did not consider the anomaly detector and only looked at the attacker's capabilities to estimate the controller's state. We showed that if and only if the controller has poles inside and on the unit circle, the attacker is able to obtain a perfect estimate of the controller's state. Furthermore, we determined when an attack is able to use a non-optimal observer, for example a Luenberger observer, to

estimate the controller state. We then classified all observer gains for the attacker, for which it can obtain a perfect estimate. Further, we proposed a defense mechanism to make it impossible for the attacker to obtain a perfect estimate. This defense mechanism is based on the introduction of uncertainty into the controller and shows similarities to watermarking approaches in the literature. Finally, we verified our results with simulations. These simulations showed that having controller poles on the unit circle might significantly slow down the attacker's estimation.

In Chapter 5, we built upon the previous chapter. Now that we know when the worst-case attack is possible, this chapter investigated how the attacker could increase its model knowledge further. More specifically, we examined how an attacker could obtain an accurate estimate of the detector's internal state. Having an estimate of the detector's state helps the attacker to design a more powerful attack. In this chapter, we focused on detectors with linear dynamics. We showed how an attacker could use a virtual detector to design an attack sequence that will remain undetected while the detector state is estimated. The attack sequence simultaneously mimics the statistics of the detector output and increases the estimation error of the operator at each time step. We utilized the Kullback-Leibler divergence as well as the concept of dual norms to find this possible attack strategy. The attack strategy was verified by applying it to the use-case of a tall wind-excited building, which is equipped with a MEWMA detector.

Based upon the previous two chapters, we know that the worst-case attack is actually feasible for certain plants. Therefore, we compared the performance of detectors in Chapter 6. The first part compared the χ^2, CUSUM, and MEWMA detectors with the metric proposed in [25]. The metric requires the worst-case impact under each detector and we proved for an attack objective based on the infinity norm that we can compute it by solving multiple convex optimization problems. The comparison of the three investigated detectors shows us that detectors with internal dynamics have the ability to limit the attack impact more than static detectors. However, the attack impact depends on the choice of the detector's parametrization, for example its forgetting factor. A problem with this detector comparison is that it depends on the attacker's objective, which is typically unknown to the operator. Therefore, we proposed a new metric in the second part of this chapter, which does not depend on the attacker's objective. We used the general detector model to show that there exists a time-invariant set, such that the attack does not trigger an alarm if it remains in this set. The size of the set is then used to compare the detectors. Comparing the detectors with this metric gives us results similar to the ones that we obtained in the first part of the chapter, but has the benefit that neither knowledge about the attacker's objective nor the plant dynamics is needed. However, the disadvantage is that the attack model we use for this metric is less powerful than the one of the worst-case attack.

Chapter 7 discussed the optimal tuning choice of an anomaly detector. The contents of this chapter can be seen as the step after the operator has chosen the preferred detector according to some metric. We proposed a Stackelberg game framework, where the defender makes the first move by picking a detector threshold.

The attacker will then design its attack in the second round of the game to maximize its impact. How the defender chooses the threshold depends on a cost function, which includes the cost for false alarms and the cost induced by the attack. We showed that there always exists a solution to this game and presented two sufficient conditions for the uniqueness of the solution. Finally, we verified the framework by applying it to a quadruple tank system equipped with a χ^2 detector.

8.2 Future Work

Let us now discuss several directions for future work.

8.2.1 Sensor Attacks

In this thesis, we considered a stealthy attack using all sensors of the closed-loop system. Since the attacker had full model knowledge as well, this is a very powerful attack model. Therefore, future work includes to look into less powerful sensor attackers. An attacker with partial sensor information is of special interest, because this kind of attacker might not be able to guarantee to not trigger an alarm due to the measurement noise. Therefore, new ways to define the attacker's stealthiness should be investigated. One example could be to consider bounded noise processes, since the noise in real word applications is probably neither Gaussian nor unbounded.

Further, we only showed that the attacker is able to break the confidentiality of a detector with linear dynamics. Hence, it would be interesting to see if the attacker could break the confidentiality of other detectors as well.

8.2.2 Actuator Attacks

This thesis only concerned sensor attacks, so a logical direction is to investigate actuator attacks as well. Since actuator attacks can have an immediate physical impact, these attacks are more dangerous to the plant's safety than sensor attacks. One direction is to see if an attacker could estimate the controller and/or detector state when it has access to all actuator signals. Further, the combination of sensor and actuator attacks is of interest. The number of necessary actuator signals and sensor measurements to break the confidentiality of the controller and/or detector should be examined. The problem of stealthy partial sensor and actuator attacks should be considered as well. Here, we also need to find a definition of what stealthiness means for this kind of attack.

8.2.3 Detector Metrics and Comparison

In the area of detector metrics and comparison we can also find many open problems. We compared the χ^2, CUSUM, and MEWMA detectors only under the assumption that the attacker uses the infinity norm to define its objective. Since the metric is impact dependent, we should also compare these detectors under different attack

objectives. The challenge in that is that it is difficult to find the worst-case impact when the problem is non-convex.

Our proposed metric in Chapter 6 is only applicable in the case of a sensor attack. Therefore, it would be interesting to see if such metrics also exists for actuator-only attacks, and attacks on both the actuators and sensors. In the case of full actuator attacks and a system with unstable zero dynamics, it is clear that the attack impact is unbounded no matter which detector is used.

In this thesis, we only looked into three different detectors, while many other detectors fit our general detector model. Hence, future work includes the comparison of other detectors. Further, it might be good to come up with novel detectors, which mitigate the attack impact significantly even for high mean times between false alarms.

Since operators have access to lots of data from their processes during nominal behavior, we could also think of using machine learning based detectors for better attack detection.

8.2.4 Experimental Validation

We showed that the worst-case attack is feasible in theory and verified our results with simulations. However, we it would be interesting to also validate the results in a real world scenario to show that these results are not only of a theoretical nature.

The need for experimental validation starts already in Chapter 3, where the detector tuning is based on the assumption of Gaussian noise. The measurement and process noise of a real process will not be Gaussian. It is therefore important to validate the mean time between false alarms, because the metric for detector comparison depends on these. Similarly, we should determine the impact the attack has on a real system as well, because it might be quite different from the impact we estimated from the linear model. This would also change our detector metric results.

The results of Chapter 4 are based on the fact that the attacker knows the exact system model. However, in real life not even an operator knows the exact model. Therefore, it would be interesting to see if the estimation of the controller's state is actually possible without the exact knowledge of the closed-loop system.

Further, mimicking the detector output statistics, as presented in Chapter 5, can fail for a real process because the statistics of the detector output are not known to the attacker.

Bibliography

[1] M. Roser. Life expectancy. (accessed: 9th of September 2019). [Online]. Available: https://ourworldindata.org/life-expectancy#life-expectancy-has-improved-globally

[2] Bundesministerium des Inneren. (2009, June) National strategy for critical infrastructure protection (CIP strategy). (accessed: 22nd of August 2019). [Online]. Available: https://www.bmi.bund.de/SharedDocs/downloads/EN/publikationen/2009/kritis_englisch.html

[3] K. Hemsley and R. Fisher, "A history of cyber incidents and threats involving industrial control systems," in *Critical Infrastructure Protection XII*, J. Staggs and S. Shenoi, Eds. Cham: Springer International Publishing, 2018, pp. 215–242.

[4] J. Slay and M. Miller, "Lessons learned from the Maroochy water breach," in *Critical Infrastructure Protection*, E. Goetz and S. Shenoi, Eds. Boston, MA: Springer US, 2008, pp. 73–82.

[5] M. Abrams and J. Weiss, *Malicious control system cyber security attack case study–Maroochy Water Services, Australia.* The MITRE corporation, 2008.

[6] R. M. Lee, M. J. Assante, and T. Conway, "Analysis of the cyber attack on the Ukrainian power grid. defense use case," *E-ISAC*, 2016.

[7] K. Zetter. (2016, March) Inside the cunning, unprecedented hack of ukraine's power grid. (accessed: 16th of Aufgust 2019). [Online]. Available: https://www.wired.com/2016/03/inside-cunning-unprecedented-hack-ukraines-power-grid/

[8] R. Langner, "Stuxnet: Dissecting a cyberwarfare weapon," *IEEE Security Privacy*, vol. 9, no. 3, pp. 49–51, May 2011.

[9] A. Greenberg. (2015, July) Hackers remotely kill a jeep on the highway - with me in it. (accessed: 24th of June 2019). [Online]. Available: https://www.wired.com/2015/07/hackers-remotely-kill-jeep-highway/

[10] ——. (2017, June) Researchers found they could hack entire wind farms. (accessed: 16th of August 2019). [Online]. Available: https://www.wired.com/story/wind-turbine-hack/?verso=true

[11] J. Leyden. (2008, January) Polish teen derails tram after hacking train network. (accessed: 16th of Aufgust 2019). [Online]. Available: https://www.theregister.co.uk/2008/01/11/tram_hack/

[12] A. Cardenas, S. Amin, B. Sinopoli, A. Giani, A. Perrig, and S. Sastry, "Challenges for securing cyber physical systems," in *Workshop on Future Directions in Cyber-physical Systems Security*. DHS, July 2009.

[13] Myndigheten för samhällsskydd och beredskap. (2014) Action plan for the protection of vital societal functions & critical infrastructure. (accessed: 7th of September 2019). [Online]. Available: https://www.msb.se/sv/publikationer/action-plan-for-the-protection-of-vital-societal-functions--critical-infrastructure/

[14] Homeland Security. (2015, September) Homeland security presidential directive 7: Critical infrastructure identification, prioritization, and protection. (accessed: 22nd of August 2019). [Online]. Available: https://www.dhs.gov/homeland-security-presidential-directive-7

[15] A. A. Cárdenas, S. Amin, and S. Sastry, "Research challenges for the security of control systems," in *Proceedings of the 3rd Conference on Hot Topics in Security*, ser. HOTSEC'08. Berkeley, CA, USA: USENIX Association, 2008, pp. 6:1–6:6.

[16] I. Hwang, S. Kim, Y. Kim, and C. E. Seah, "A survey of fault detection, isolation, and reconfiguration methods," *IEEE Transactions on Control Systems Technology*, vol. 18, no. 3, pp. 636–653, May 2010.

[17] S. H. Kafash, J. Giraldo, C. Murguia, A. A. Cardenas, and J. Ruths, "Constraining attacker capabilities through actuator saturation," in *2018 Annual American Control Conference (ACC)*, June 2018, pp. 986–991.

[18] C. Murguia and J. Ruths, "On reachable sets of hidden cps sensor attacks," in *2018 Annual American Control Conference (ACC)*, June 2018, pp. 178–184.

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

[20] P. Hespanhol, M. Porter, R. Vasudevan, and A. Aswani, "Dynamic watermarking for general LTI systems," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, Dec 2017, pp. 1834–1839.

[21] B. Satchidanandan and P. R. Kumar, "Dynamic watermarking: Active defense of networked cyber–physical systems," *Proceedings of the IEEE*, vol. 105, no. 2, pp. 219–240, Feb 2017.

[22] Y. Mo and B. Sinopoli, "False data injection attacks in control systems," in *First Workshop on Secure Control Systems*, April 2010.

[23] Z. Guo, D. Shi, K. H. Johansson, and L. Shi, "Worst-case stealthy innovation-based linear attack on remote state estimation," *Automatica*, vol. 89, pp. 117 – 124, 2018.

[24] A. A. Cárdenas, S. Amin, Z. Lin, Y. Huang, C. Huang, and S. Sastry, "Attacks against process control systems: Risk assessment, detection, and response," in *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*, ser. ASIACCS '11. New York, NY, USA: ACM, 2011, pp. 355–366.

[25] D. I. Urbina, J. A. Giraldo, A. A. Cárdenas, N. O. Tippenhauer, J. Valente, M. Faisal, J. Ruths, R. Candell, and H. Sandberg, "Limiting the impact of stealthy attacks on industrial control systems," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '16. New York, NY, USA: ACM, 2016, pp. 1092–1105.

[26] A. Ghafouri, W. Abbas, A. Laszka, Y. Vorobeychik, and X. Koutsoukos, "Optimal thresholds for anomaly-based intrusion detection in dynamical environments," in *Decision and Game Theory for Security*. Springer International Publishing, 2016, pp. 415–434.

[27] Y. Z. Lun, A. D'Innocenzo, F. Smarra, I. Malavolta, and M. D. D. Benedetto, "State of the art of cyber-physical systems security: An automatic control perspective," *Journal of Systems and Software*, vol. 149, pp. 174 – 216, 2019.

[28] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "A secure control framework for resource-limited adversaries," *Automatica*, vol. 51, pp. 135 – 148, 2015.

[29] F. Pasqualetti, F. Dörfler, and F. Bullo, "Control-theoretic methods for cyber-physical security: Geometric principles for optimal cross-layer resilient control systems," *IEEE Control Systems*, vol. 35, no. 1, pp. 110–127, 2015.

[30] A. Duz, S. Phillips, A. Fagiolini, R. G. Sanfelice, and F. Pasqualetti, "Stealthy attacks in cloud-connected linear impulsive systems," in *2018 Annual American Control Conference (ACC)*, June 2018, pp. 146–152.

[31] R. S. Smith, "Covert misapproriation of networked control systems: Presenting a feedback structure," *IEEE Control Systems*, vol. 35, no. 1, pp. 82–92, 2015.

[32] G. Park, H. Shim, C. Lee, Y. Eun, and K. H. Johansson, "When adversary encounters uncertain cyber-physical systems: Robust zero-dynamics attack with disclosure resources," in *2016 IEEE 55th Conference on Decision and Control (CDC)*, Dec 2016, pp. 5085–5090.

[33] T. Lipp and S. Boyd, "Antagonistic control," *Systems & Control Letters*, vol. 98, pp. 44 – 48, 2016.

[34] Y. Mo, S. Weerakkody, and B. Sinopoli, "Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs," *IEEE Control Systems*, vol. 35, no. 1, pp. 93–109, Feb 2015.

[35] S. Amin, A. A. Cárdenas, and S. S. Sastry, "Safe and secure networked control systems under denial-of-service attacks," in *Hybrid Systems: Computation and Control*, R. PMajumdar and P. Tabuada, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 31–45.

[36] V. Ugrinovskii and C. Langbort, "Controller–jammer game models of denial of service in control systems operating over packet-dropping links," *Automatica*, vol. 84, pp. 128 – 141, 2017.

[37] V. S. Dolk, P. Tesi, C. De Persis, and W. P. M. H. Heemels, "Event-triggered control systems under denial-of-service attacks," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 1, pp. 93–105, March 2017.

[38] M. Xue, W. Wang, and S. Roy, "Security concepts for the dynamics of autonomous vehicle networks," *Automatica*, vol. 50, no. 3, pp. 852 – 857, 2014.

[39] Y. Yuan and Y. Mo, "Security in cyber-physical systems: Controller design against known-plaintext attack," in *2015 54th IEEE Conference on Decision and Control (CDC)*, Dec 2015, pp. 5814–5819.

[40] S. M. Dibaji, M. Pirani, A. M. Annaswamy, K. H. Johansson, and A. Chakrabortty, "Secure control of wide-area power systems: Confidentiality and integrity threats," in *2018 IEEE Conference on Decision and Control (CDC)*, Dec 2018, pp. 7269–7274.

[41] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Transactions on Automatic Control*, vol. 59, no. 6, pp. 1454–1467, June 2014.

[42] M. S. Chong, M. Wakaiki, and J. P. Hespanha, "Observability of linear systems under adversarial attacks," in *2015 American Control Conference (ACC)*, July 2015, pp. 2439–2444.

[43] J. M. Hendrickx, K. H. Johansson, R. M. Jungers, H. Sandberg, and K. C. Sou, "Efficient computations of a security index for false data attacks in power networks," *IEEE Transactions on Automatic Control*, vol. 59, no. 12, pp. 3194–3208, Dec 2014.

[44] J. Milošević, H. Sandberg, and K. H. Johansson, "A security index for actuators based on perfect undetectability: Properties and approximation," in *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Oct 2018, pp. 235–241.

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

[46] H. Sandberg and A. M. H. Teixeira, "From control system security indices to attack identifiability," in *2016 Science of Security for Cyber-Physical Systems Workshop (SOSCYPS)*, April 2016, pp. 1–6.

[47] J. Milošević, A. Teixeira, T. Tanaka, K. H. Johansson, and H. Sandberg, "Security measure allocation for industrial control systems: Exploiting systematic search techniques and submodularity," *International Journal of Robust and Nonlinear Control*, 2018.

[48] I. Jovanov and M. Pajic, "Sporadic data integrity for secure state estimation," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, Dec 2017, pp. 163–169.

[49] ——, "Secure state estimation with cumulative message authentication," in *2018 IEEE Conference on Decision and Control (CDC)*, Dec 2018, pp. 2074–2079.

[50] P. T. Devanbu and S. Stubblebine, "Software engineering for security: A roadmap," in *Proceedings of the Conference on The Future of Software Engineering*, ser. ICSE '00. New York, NY, USA: ACM, 2000, pp. 227–239.

[51] R. M. G. Ferrari and A. M. H. Teixeira, "Detection and isolation of routing attacks through sensor watermarking," in *2017 American Control Conference (ACC)*, May 2017, pp. 5436–5442.

[52] R. M. Ferrari and A. M. Teixeira, "Detection and isolation of replay attacks through sensor watermarking," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 7363 – 7368, 2017, 20th IFAC World Congress.

[53] A. M. H. Teixeira and R. M. G. Ferrari, "Detection of sensor data injection attacks with multiplicative watermarking," in *2018 European Control Conference (ECC)*, June 2018, pp. 338–343.

[54] J. Tian, R. Tan, X. Guan, and T. Liu, "Enhanced hidden moving target defense in smart grids," *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 2208–2223, March 2019.

[55] J. Giraldo, A. Cardenas, and R. G. Sanfelice, "A moving target defense to detect stealthy attacks in cyber-physical systems," in *2019 American Control Conference (ACC)*, July 2019, pp. 391–396.

[56] K. Kogiso and T. Fujita, "Cyber-security enhancement of networked control systems using homomorphic encryption," in *2015 54th IEEE Conference on Decision and Control (CDC)*, Dec 2015, pp. 6836–6843.

[57] F. Farokhi, I. Shames, and N. Batterham, "Secure and private control using semi-homomorphic encryption," *Control Engineering Practice*, vol. 67, pp. 13 – 20, 2017.

[58] M. H. Manshaei, Q. Zhu, T. Alpcan, T. Bacşar, and J.-P. Hubaux, "Game theory meets network security and privacy," *ACM Comput. Surv.*, vol. 45, no. 3, pp. 25:1–25:39, Jul. 2013.

[59] J. Pita, M. Jain, F. Ordóñez, C. Portway, M. Tambe, C. Western, P. Paruchuri, and S. Kraus, "Using game theory for Los Angeles airport security," *AI Magazine*, vol. 30, no. 1, p. 43, Jan. 2009.

[60] S. R. Etesami and T. Başar, "Dynamic games in cyber-physical security: An overview," *Dynamic Games and Applications*, Jan 2019.

[61] Q. Zhu and T. Başar, "Game-theoretic methods for robustness, security, and resilience of cyberphysical control systems: Games-in-games principle for optimal cross-layer resilient control systems," *IEEE Control Systems*, vol. 35, no. 1, pp. 46–65, Feb 2015.

[62] F. Miao, M. Pajic, and G. J. Pappas, "Stochastic game approach for replay attack detection," in *52nd IEEE Conference on Decision and Control*, Dec 2013, pp. 1854–1859.

[63] P. Shukla, A. Chakrabortty, and A. Duel-Hallen, "A cyber-security investment game for networked control systems," in *2019 American Control Conference (ACC)*, July 2019, pp. 2297–2302.

[64] M. O. Sayin and T. Başar, "Secure sensor design for cyber-physical systems against advanced persistent threats," in *Decision and Game Theory for Security*, S. Rass, B. An, C. Kiekintveld, F. Fang, and S. Schauer, Eds. Cham: Springer International Publishing, 2017, pp. 91–111.

[65] K. Chen, V. Gupta, and Y. F. Huang, "Minimum variance unbiased estimation in the presence of an adversary," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, Dec 2017, pp. 151–156.

[66] L. Niu and A. Clark, "Secure control under linear temporal logic constraints," in *2018 Annual American Control Conference (ACC)*, June 2018, pp. 3544–3551.

[67] A. Tsiamis, K. Gatsis, and G. J. Pappas, "State-secrecy codes for stable systems," in *2018 Annual American Control Conference (ACC)*, June 2018, pp. 171–177.

[68] ——, "State estimation codes for perfect secrecy," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, Dec 2017, pp. 176–181.

[69] Y. Wang, Z. Huang, S. Mitra, and G. E. Dullerud, "Differential privacy in linear distributed control systems: Entropy minimizing mechanisms and performance tradeoffs," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 1, pp. 118–130, March 2017.

[70] J. Giraldo, A. A. Cardenas, and M. Kantarcioglu, "Security vs. privacy: How integrity attacks can be masked by the noise of differential privacy," in *2017 American Control Conference (ACC)*, May 2017, pp. 1679–1684.

[71] E. Nekouei, T. Tanaka, M. Skoglund, and K. H. Johansson, "Information-theoretic approaches to privacy in estimation and control," *Annual Reviews in Control*, vol. 47, pp. 412 – 422, 2019.

[72] E. Nekouei, M. Skoglund, and K. H. Johansson, "Privacy of information sharing schemes in a cloud-based multi-sensor estimation problem," in *2018 Annual American Control Conference (ACC)*, June 2018, pp. 998–1002.

[73] C. A. Lowry, W. H. Woodall, C. W. Champ, and S. E. Rigdon, "A multivariate exponentially weighted moving average control chart," *Technometrics*, vol. 34, no. 1, pp. 46–53, 1992.

[74] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954.

[75] C. D. Brown and H. T. Davis, "Receiver operating characteristics curves and related decision measures: A tutorial," *Chemometrics and Intelligent Laboratory Systems*, vol. 80, no. 1, pp. 24 – 38, 2006.

[76] B. M. Adams, W. H. Woodall, and C. A. Lowry, "The use (and misuse) of false alarm probabilities in control chart design," in *Frontiers in Statistical Quality Control 4*, H.-J. Lenz, G. B. Wetherill, and P.-T. Wilrich, Eds. Heidelberg: Physica-Verlag HD, 1992, pp. 155–168.

[77] T. Tarn and Y. Rasis, "Observers for nonlinear stochastic systems," *IEEE Transactions on Automatic Control*, vol. 21, no. 4, pp. 441–448, August 1976.

[78] G. C. Runger and S. S. Prabhu, "A Markov chain model for the multivariate exponentially weighted moving averages control chart," *Journal of the American Statistical Association*, vol. 91, no. 436, pp. 1701–1706, 1996.

[79] K. Zhou and J. C. Doyle, *Essentials of Robust Control*. Prentice-Hall, 1999.

[80] S. W. Chan, G. Goodwin, and K. S. Sin, "Convergence properties of the Riccati difference equation in optimal filtering of nonstabilizable systems," *IEEE Transactions on Automatic Control*, vol. 29, no. 2, pp. 110–118, February 1984.

[81] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*. Englewood Cliffs, NJ: Prentice-Hall, 1979.

[82] C. de Souza, M. Gevers, and G. Goodwin, "Riccati equations in optimal filtering of nonstabilizable systems having singular state transition matrices," *IEEE Transactions on Automatic Control*, vol. 31, no. 9, pp. 831–838, Sep. 1986.

[83] J. C. Doyle, B. A. Francis, and A. R. Tannenbaum, *Feedback control theory*. Courier Corporation, 2013.

[84] C. Johnson, "Accomodation of external disturbances in linear regulator and servomechanism problems," *IEEE Transactions on Automatic Control*, vol. 16, no. 6, pp. 635–644, December 1971.

[85] J. Chen and C. N. Nett, "Sensitivity integrals for multivariable discrete-time systems," *Automatica*, vol. 31, no. 8, pp. 1113 – 1124, 1995.

[86] G. Stein, "Respect the unstable," *IEEE Control Systems Magazine*, vol. 23, no. 4, pp. 12–25, Aug 2003.

[87] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[88] D. G. Luenberger, *Optimization by Vector Space Methods*, 1st ed. New York, NY, USA: John Wiley & Sons, Inc., 1997.

[89] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. New York, NY, USA: Cambridge University Press, 2012.

[90] J. N. Yang, A. K. Agrawal, B. Samali, and J.-C. Wu, "Benchmark problem for response control of wind-excited tall buildings," *Journal of Engineering Mechanics*, vol. 130, no. 4, pp. 437–446, 2004.

[91] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[92] K. H. Johansson, "The quadruple-tank process: a multivariable laboratory process with an adjustable zero," *IEEE Transactions on Control Systems Technology*, vol. 8, no. 3, pp. 456–465, May 2000.

[93] O. Güler, *Foundations of Optimization*. Springer-Verlag New York, 2010.

[94] K. Leyton-Brown and Y. Shoham, "Essentials of game theory: A concise multidisciplinary introduction," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 2, no. 1, pp. 1–88, 2008.

[95] G. Leitmann, "On generalized Stackelberg strategies," *Journal of Optimization Theory and Applications*, vol. 26, no. 4, pp. 637–643, Dec 1978.

[96] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. McGraw-Hill, 1976.