DEGREE PROJECT IN ELECTRICAL ENGINEERING,
SECOND CYCLE, 30 CREDITS
STOCKHOLM, SWEDEN 2016

Security Analysis of Control System Anomaly Detectors

DAVID UMSONST

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ELECTRICAL ENGINEERING

Abstract

Anomaly detectors in control systems are used to detect system faults and they are typically based on an analytical system model, which generates residual signals to find a fault. The detectors are designed to detect randomly occurring faults but not coordinated malicious attacks on the system.
Therefore three different anomaly detectors, namely a detector solely based on the last residual, a multivariate exponentially weighted moving average filter and a cumulative sum, are investigated to determine which detector yields the smallest worst-case impact of a time-limited data injection attack.
For this reason optimal control problems are formulated to characterize the worst-case attack under different anomaly detectors, which lead to non-convex optimization problems. Relaxations to convex problems are proposed and solved numerically and in special cases also analytically.
The detectors are compared by solving the optimal control problems for a simple simulation example as well as a quadruple-tank process. Simulations and experiments show that the cumulative sum seems to be the detector to choose, if one wants to limit the worst-case attack impact.

Abstract

Anomalidetektorer i styrsystem används normalt för att detektera systemfel och de är oftast baserade på en analytisk systemmodell vilken genererar residualsignaler för att upptäcka felen. Detektorerna är oftast konstruerade för att upptäcka slumpmässigt förekommande fel och inte samordnade angrepp på systemet.
Därför utvärderas här tre olika anomalidetektorer: en detektor som enbart grundar sig på den senaste residualen, en som är baserad på multivariat exponentiellt viktat glidande medelärde och en kumulativ summa. I utvärderingen undersöker vi vilken detektor som mest begränsar en attack i form av en datainjektion.
Av denna anledning formuleras optimala styrproblem för att karakterisera den värsta attacken för de olika anomalidetektorerna, vilket leder till ickekonvexa optimeringsproblem. Relaxeringar till konvexa problem föreslås och löses numeriskt och i särskilda fall även analytiskt. Detektorerna jämförs genom att lösa de optimala styrproblem för ett simuleringsexempel såväl som för en riktig fyrtanksprocess. Både simuleringar och experiment visar att den kumulativa summan är den detektor som begränsar de studerade attackerna mest.

Contents

List of Symbols and Abbreviations

List of Symbols

The list of symbols contains the most important symbols used in the thesis. We neglect the unit of the symbols, because they mostly depend on the process used and can not be stated generally.

Symbol	Description
k	Discrete time variable
x_k	System state
u_k	Actuator signal
$\tilde{u}_k$	Corrupted actuator signal
y_k	Measurement signal
$\tilde{y}_k$	Corrupted measurement signal
z_k	Observer state
r_k	Residual signal
w_k	Process noise
v_k	Measurement noise
a_k	Attack signal
a_k^f	Physical attack
a_k^u	Actuator attack
a_k^y	Measurement attack
A	System matrix
B	Control input matrix
C	Measurement matrix
K	Control matrix
L	Observer matrix
B_a	Attack input matrix
D_a	Attack measurement matrix
μ_k	Extended system state
A_e	Extended system matrix
B_e	Extended input matrix
C_e	Extended measurement matrix
D_e	Extended feed-through matrix
N	Duration of the attack
ΔN	Time steps the attacker considers after the attack happened
S_k	Security metric of the anomaly detector

J_D	Threshold of the anomaly detector
δ	Forgetting factor of the cumulative sum
β	Forgetting factor of the exponentially weighted moving average filter
$\mu_{k,s}$	Extended system state with linear anomaly detector included
$r_{k,s}$	Residual of the extended system state with linear anomaly detector included
$A_{e,s}$	Extended system matrix with linear anomaly detector included
$B_{e,s}$	Extended input matrix with linear anomaly detector included
$C_{e,s}$	Extended measurement matrix with linear anomaly detector included
$D_{e,s}$	Extended feed-through matrix with linear anomaly detector included
J_{th}	Threshold of the stateless detector
e_k	Difference between system and observer state
Σ_{r_k}	Covariance matrix of the residual signal
Σ_{e_k}	Covariance matrix of the error signal
Σ_{w_k}	Covariance matrix of the process noise
Σ_{v_k}	Covariance matrix of the measurement signal
μ	Trajectory of the extended system from $k = 1$ to $k = N$
a	Trajectory of the attack signals from $k = 0$ to $k = N - 1$
r	Trajectory of the residual signals from $k = 0$ to $k = N - 1$
x	Trajectory of the system states from $k = 1$ to $k = N$
T_x	Matrix that extracts x from μ
A_c	Extended system matrix over the attack horizon
B_c	Extended input matrix over the attack horizon
C_c	Extended measurement matrix over the attack horizon
D_c	Extended feed-through matrix over the attack horizon
c_a	Upper bound for the attack signal

Abbreviations

Abbreviation	**Complete Description**
CUSUM	Cumulative Sum
EWMA	Exponentially Weighted Moving Average Filter
MEWMA	Multivariate Exponentially Weighted Moving Average Filter
IT	Information Technology

Chapter 1

Introduction

In recent years a great interest in cyber-physical security based on control theory has developed. A problem with security measures purely based on information technology (IT) is that these security measurements do not consider the physical but only the cyber side of the cyber-physical systems to protect, so it protects only the data sent over the network without checking if the values physically make sense. For example, if a signal is changed with malice aforethought before it is sent through the network the IT-based protection will not see anything wrong with the data.
Therefore, security measures based on control-theoretic results are becoming more and more popular, since they use analytical models of the cyber-physical systems to check if the signal received actually behaves according to the physics behind the system. These control-theoretic measures are by no means a replacement for the IT-based security measurement but an additional layer of protection to secure the system controlled.
In control theory, anomaly detectors are used to detect odd behavior and these are often designed to detect randomly occurring faults in the system but not malicious attacks. Hence, an intelligent adversary is capable of creating an attack, which deteriorates the systems performance while remaining undetected.
The most commonly used detector in recent papers is based only on the current residual signal, which is determined by comparing actual measurements with predicted measurement signals. Our work investigates and compares this and two more anomaly detectors. The other two detectors are a cumulative sum and a multivariate exponentially weighted moving average filter, that consider past measurement and control signals as well. Hence, they have a memory of past events, while the commonly used detector only considers the present event.
To compare the detectors we characterize stealthy worst-case attacks and analyze the impact of these attacks under the different detectors. The worst-case impacts are obtained by formulating the attack scenario as an optimization problem, which maximizes the impact on the system and remains undetected by the detector at the same time. This leads to non-convex optimization problems, so relaxations to convex problems are proposed, which are then solved numerically and in special cases also analytically.
These worst-case attacks are analyzed for different detector configurations using a simple simulation as well as a more sophisticated quadruple tank process model. Furthermore an experiment on a real quadruple tank process is con-

ducted to see how realistic the worst-case attacks are.
We observe that the cumulative sum detector restricts the adversary in a better way compared to the other two detectors, because it decreases the attack impact compared to the stateless detector. The multivariate exponentially weighted moving average detector can actually benefit the attacker in certain scenarios investigated.

1.1 Related Work

Due to the growing interest in cyber-physical security in recent times many articles and papers are published and a few of them are reviewed here.
In [2] an attack space for the attacker is defined depending on the disclosure resources, disruption resources, and the model knowledge of the attacker. Several attacks are fit into the attack space such as, the replay attack, which replays recorded data while it is attacking, and the zero dynamics attack, which takes advantage of the system's zero dynamics. Furthermore the novel bias injection attack is introduced in this article, which injects a bias to the steady state of the system without being detected. The reference [3] aims for a more general approach to define attacks as well. The networked control systems are described as descriptor systems and several attacks can be fit into this scheme, e.g. false data injection attacks or covert attacks. Furthermore a definition for the detectability and identifiability of attacks is presented in this article.
While [2] and [3] focus on several attack scenarios other articles concentrate mainly on one attack scheme and propose detection mechanisms, see [4], for example. The replay attack is investigated in [4] and watermarking of the control inputs is used to detect the attack. The idea is to add noise with known time varying statistics to the control inputs as a watermarking so that the anomaly detector can check for these statistics to see if the correct measurements are received.
Another paper which focuses mainly on one attack is [5], which explores the covert attack. A linear and a nonlinear structure for a covert attack are presented. For the linear covert attack several beneficial factors for its undetectability are exposed, e.g. the model error from the attackers model knowledge to the actual model, but also that a sophisticated controller in the system can help an attacker to stay undetected. The nonlinear covert attack is deemed to be more difficult to implement than the linear one due to the changes in the operating points of the system caused by the attack.
Hendrickx et al. [6] has again a different approach of handling the threat of an attack. This article tries to identify weaknesses in a static model of a power grid by calculating a security index for each measurement taken. The security index for a measurement indicates how many measurements have to be altered to create a undetectable attack on this particular measurement. The security index for all measurements shows then which measurements one should protect to make it harder for an adversary to attack the system. A fast algorithm for computing the security indices is proposed, which provides either the exact solution or an approximation with an upper bound on the error.
Teixeira et al. [7] proposes risk management strategies to deal with the threat of attacks, where attacks are categorized in a risk management context according to their likelihood and impact. This can be used to determine the threat an

attack poses. In this context minimum-resource attacks are defined, which are closely related to the security index problem in [6]. Furthermore a minimum-resource, maximum-impact attack for dynamic systems is defined, which can also be used in the risk management context to determine the likelihood and impact of an attack.
Teixeira et al. [8] proposes a new metric, which tells how sensitive the system is to stealthy false-data injection attacks. The so called output-to-output l_2-gain is defined as the decrease in the systems performance an attack can cause under the condition of remaining stealthy. Moreover necessary and sufficient conditions for the output-to-output l_2-gain to be bounded are given, which are closely related to the zero dynamics of the system.
Urbina et al. [1] present a comprehensive survey on the current literature about security of cyber-physical systems and put them in a unified taxonomy to compare the literature. The taxonomy is based on the system model, a trust model, that states which system components are trusted, the detector used and how the detector's performance is evaluated. Furthermore a new metric based on the mean time between false-alarms and the attack impact is presented to compare different anomaly detectors. Experiments and simulations show that a cumulative sum detector performs better than a stateless detector according to the new metric.

1.2 Outline

Firstly we introduce the necessary background information to follow the course of our work in Chapter 2, including the background on networked control systems, the attack model used and the anomaly detectors investigated.
The theoretical results are presented in Chapter 3, where the worst-case attacks under different detectors are characterized as optimal control problems. Furthermore reformulations and solutions to the optimization problems are proposed.
The simulation and experimental results on the simple simulation example and the quadruple tank process are shown in Chapter 4. Chapter 5 concludes our work with a discussion of the experimental and simulation results and an overall conclusion with future work suggestions.

Chapter 2

Background

This chapter presents the necessary background needed to follow the course of our work. Starting with a general introduction to networked control systems the section continues with the description of the control and adversary model and concludes with the anomaly detectors considered.

2.1 Background on Networked Control Systems

At first some general background information on network control systems is given.
In a networked control system the components are not directly connected and information, like sensor measurements and control signals, are sent over a communication network (see Figure 2.1). Examples for these kind of systems are electrical power networks, where the networks data is gathered in a control center and control signals are sent out to all plants, or industrial processes.

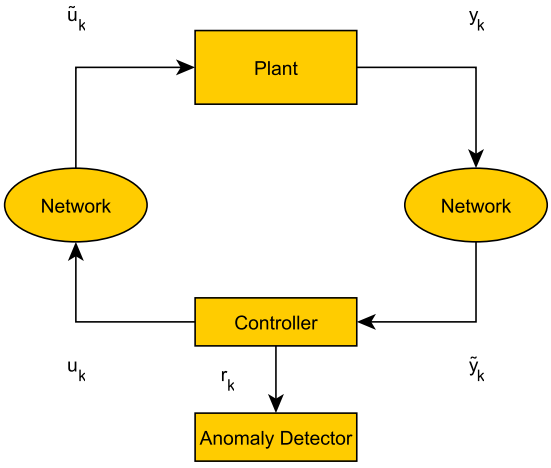

Figure 2.1: Block Diagram of a Networked Control System

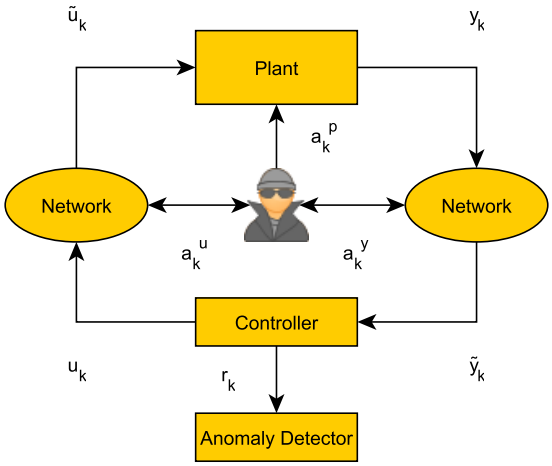

Figure 2.2: Block Diagram of a Networked Control System under Attack

The basic components of a networked control system are

- The physical plant, which has to be controlled to reach a desired performance.
- The controller, which determines the control input u_k according to the measurements $\tilde{y}_k$ to get the desired performance.
- The network, which is used to exchange data between the controller and the plant. Note that the information going through the network can be different before (u_k, y_k) and after ($\tilde{u}_k$, $\tilde{y}_k$) passing the network, due to noise, package losses or malicious changes of an attacker, for example.
- The anomaly detector, which is used to detect abnormal system behavior like faults by using measurements and control signals.

This structure makes the system vulnerable to attacks, where an adversary places itself in the middle of the control loop (see Figure 2.2). This adversary can eavesdrop on the signals send through the network, manipulate them or do both. Furthermore the adversary might have direct access to the plant to change the system directly. These abilities can be described as disclosure and disruption resources and with the model knowledge of the adversary they span the attack space presented in [2] (see Figure 2.3). Several attacks can be classified in this attack space, like the Denial-of-Service attack or a zero dynamics attack, where the attacker makes use of the system's zero dynamics. In our case the attacker has no disclosure resources and is placed in the plane spanned by the disruption resources and the model knowledge.
A practical example is a smart grid, where the adversary wants to steal electrical power. He changes the measurement data, which is sent to the control center to hide his actions. The anomaly detector should detect the wrong measurements due to the difference in the predicted and real measurement and trigger an alarm, if the attack is not stealthy.

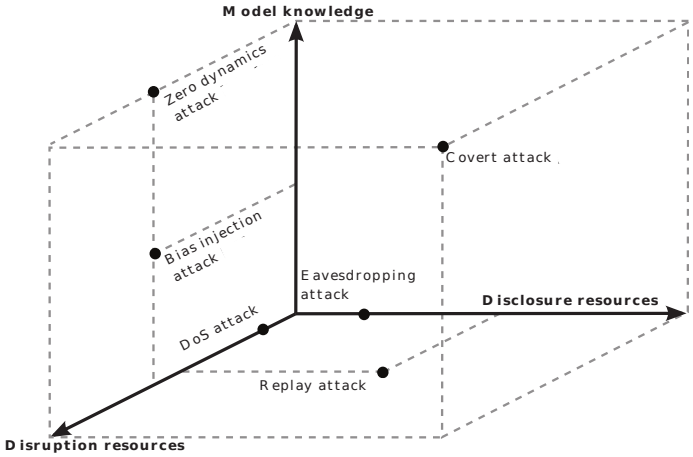

Figure 2.3: Attack Space from [2]

2.2 Modeling Networked Control Systems and Attacks

A networked control system as described above is a control system, which uses a network to exchange data like control inputs or sensor measurements. The physical plant can be modeled as a discrete-time state-space model

$$\begin{aligned} x_{k+1} &= Ax_k + B\tilde{u}_k + w_k \text{ with } x_0 \text{ given} \\ y_k &= Cx_k + v_k, \end{aligned}$$

where $x_k \in \mathbb{R}^n$ is the state of the system, $x_0 \in \mathbb{R}^n$ the initial state of the system, $\tilde{u}_k \in \mathbb{R}^l$ the control input received over the network, $y_k \in \mathbb{R}^p$ the measurements and $w_k \in \mathbb{R}^n$ and $v_k \in \mathbb{R}^p$ are the zero mean Gaussian process and measurement noise, respectively. Here $A \in \mathbb{R}^{n \times n}$ is the system matrix, $B \in \mathbb{R}^{n \times l}$ is the control input matrix and $C \in \mathbb{R}^{p \times n}$ the measurement matrix
It is assumed that the system is controlled with a state feedback controller based on a state observer

$$\begin{aligned} z_{k+1} &= Az_k + Bu_k + L(\tilde{y}_k - Cz_k) \text{ with } z_0 \text{ given} \\ r_k &= \tilde{y}_k - Cz_k \\ u_k &= -Kz_k. \end{aligned}$$

Here $z_k \in \mathbb{R}^n$ is the state of the observer, $u_k \in \mathbb{R}^l$ the calculated control input, $\tilde{y}_k \in \mathbb{R}^p$ the measurements received over the network and $r_k \in \mathbb{R}^p$ is the residual at time instance k. We assume the control matrix $K \in \mathbb{R}^{l \times n}$ and observer matrix $L \in \mathbb{R}^{n \times p}$ are chosen so that the system and error dynamics are stable.
The values of $\tilde{u}_k$ and $\tilde{y}_k$ can be different from the values of u_k and y_k due to data loss, noise in the network or due to a malicious attack $a_k \in \mathbb{R}^m$ on the system, for example. In the following we want to introduce the attack model. We can distinguish between three different ways to influence the system

1. **Physical attacks** $a_k^p \in \mathbb{R}^n$
   This attacks directly influence the state x_k by causing a physical alteration

in the plant. An example could be punching a hole into a water tank, so the tank loses water continuously.

2. **Actuator attacks** $a_k^u \in \mathbb{R}^l$
   These attacks influence the actuator signal u_k

$$\tilde{u}_k = u_k + a_k^u.$$

3. **Sensor attacks** $a_k^y \in \mathbb{R}^p$
   Sensor attacks are the equivalent to the actuator attacks but on sensors

$$\tilde{y}_k = y_k + a_k^y.$$

However one should keep in mind that the sensor and actuator attacks do not have to attack all actuator values and sensor measurements.
Stacking these attacks on top of each other leads to the attack vector

$$a_k = \begin{bmatrix} a_k^p \\ a_k^u \\ a_k^y \end{bmatrix}.$$

Introducing the attack to the plant and observer results in

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + B_a a_k + w_k \text{ with } x_0 \text{ given} \\ y_k &= Cx_k + v_k \\ z_{k+1} &= Az_k + Bu_k + L(y_k + D_a a_k + v_k - Cz_k) \text{ with } z_0 \text{ given} \\ r_k &= y_k + D_a a_k + v_k - Cz_k \\ u_k &= -Kz_k. \end{aligned}$$

Here B_a represents the influence the attack can directly have on the state by either a physical or an actuator attack and D_a the influence of the attacks on the measurements using sensor attacks. Due to the separation of the attacks in attacks on the states and the measurements the attack matrices often take the structure

$$B_a = [B_a^p, B_a^a, \mathbf{0}] \text{ and } D_a = [\mathbf{0}, \mathbf{0}, D_a^s] \tag{2.1}$$

where $\mathbf{0}$ are zero matrices of appropriate dimensions for the physical, actuator and sensor attacks, respectively. This structure has some interesting consequences later in this thesis (see Section 3.2.1).
We can combine the plant and the observer systems to get an extended system with $\mu_k = [x_k^T, z_k^T]^T$ as the extended systems state, the attack a_k as the input and the residual r_k as the system output

$$\begin{aligned} \mu_{k+1} &= A_e \mu_k + B_e a_k + \begin{bmatrix} w_k \\ Lv_k \end{bmatrix} \\ r_k &= C_e \mu_k + D_e a_k + v_k \end{aligned}$$

with

$$A_e = \begin{bmatrix} A & -BK \\ LC & A - BK - LC \end{bmatrix}, B_e = \begin{bmatrix} B_a \\ LD_a \end{bmatrix}, C_e = \begin{bmatrix} C & -C \end{bmatrix} \text{ and } D_e = D_a.$$

The initial state is given by $\mu_0 = [x_0^T, z_0^T]^T$. Since K and L are assumed to be chosen so that the plant and the error dynamics are stable, A_e is stable as well. The residual r_k is used to determine how much the real system state deviates from the estimated state given by the observer. Therefore it can be used to detect faults or attacks on the system if the residual does not stay close to zero meaning the estimated state does not coincide with the real state.
For the attacker we assume the following

- The attack duration is finite and lasts N steps, because we assume that the attacker has limited resources, which includes the time to attack as well
- After N time steps the attacker is not able to influence the system again, hence

  $$a_k = 0 \ \forall k \geq N$$

  assuming without loss of generality that the attack starts at $k = 0$.
- The attacker wants to remain stealthy, hence the attack should not trigger any alarm in the anomaly detector.
- The attacker has perfect model knowledge

The noise terms are neglected in the systems model for the sake of brevity.

2.3 Anomaly detectors

The attacks mentioned in the previous section are an unwelcome change of the extended system state and to detect their presence one can use anomaly detectors, which are actually designed to detect randomly occurring faults.
The basic principle of operation of the anomaly detectors is that at time instance l a metric S_l is calculated based on system data, e.g. control inputs u_l and measurements $\tilde{y}_l$ in control systems, and an alarm is triggered if S_l is greater than a threshold J_D. In our case S_l will be calculated using the residuals r_l and the calculation of S_l depends on the anomaly detector used. An anomaly detector is called *stateless* anomaly detector if it just considers the current system data at time $l = k$, i.e. u_k and $\tilde{y}_k$ in form of the current residual r_k, because it does not have a memory of the system's past. A *stateful* detector on the other hand considers past and presents values, u_l and $\tilde{y}_l$ with $l \leq k$ to determine the current metric S_k. Three detectors are considered and described in the following.

2.3.1 Stateless Anomaly Detector

One of the simplest anomaly detectors one could imagine is a detector solely based on the current residual r_k. The one used in our work is

$$S_{k+1} = ||r_k||_2^2.$$

The squared Euclidean norm is used because it leads to a simpler mathematical treatment later on.

2.3.2 Cumulative Sum

The cumulative sum (CUSUM) was developed by E.S. Page [9]. The CUSUM algorithm is used to detect small changes in a variable θ of the signal investigated, e.g. θ could be the mean of the signal and it changes from its initial value θ^0 to θ^1 due to an attack. CUSUM is defined as

$$S_{k+1} = \max(0, S_k + g_k) \text{ with } S_0 = 0$$

where g_k has to have a positive trend ($g_k > 0$) if a change in θ occurs and a negative trend ($g_k \leq 0$) if no change in θ occurs.
Lorden [10] proved that CUSUM is optimal in the sense that it has the smallest average delay to detect a change after it occured under the condition that the false alarm rate approaches zero asymptotically. His proof considers a parametric CUSUM algorithm meaning that we know to which value θ will change from the initial value θ^0 and that we know the probability density functions of θ before and after the change $f_{\theta^0}(x)$ and $f_{\theta^1}(x)$, respectively. Here x is the signal, in which the parameter change occurs. CUSUM is then defined as

$$S_{k+1} = \max(0, S_k + \log\left(\frac{f_{\theta^1}(x_k)}{f_{\theta^0}(x_k)}\right))$$

with $S_0 = 0$ and $g_k = \log(\frac{f_{\theta^1}(x_k)}{f_{\theta^0}(x_k)})$ is the logarithmic likelihood ratio.
Later on it was proved that CUSUM is also the optimal change detection algorithm, when the false alarm rate does not approach zero asymptotically [11] [12]. Therefore CUSUM is the optimal change detection algorithm if we want to detect a constant change from θ^0 in a minimal time while guaranteeing a certain false-alarm rate. It seems like CUSUM is the anomaly detector to use due to its optimality, but in general we do not know the change an attack will cause in the residual. Hence we have to use a non-parametric CUSUM algorithm. Similar to [1], the non-parametric CUSUM algorithm used here is

$$S_{k+1} = \max(0, S_k + ||r_k||_2^2 - \delta) \text{ with } S_0 = 0.$$

The forgetting factor δ is used to avoid too many false alarms by eliminating the naturally occurring noise in the residual. The difference between the forgetting factor δ and the threshold J_D is that the forgetting factor is used to eliminate the noise influence in the current residual while J_D is used to detect if there is significant change in the residual signal over time. The forgetting factor should be chosen so that $\mathbb{E}[||r_k||_2^2 - \delta] \leq 0$ to achieve a negative trend under no attack, but also $\delta \leq J_D$ has to be satisfied otherwise we forget more than we want to detect.

2.3.3 Exponentially Weighted Moving Average Filter

The third anomaly detector is the exponentially weighted moving average (EWMA) filter

$$S_{k+1} = \beta r_k + (1 - \beta) S_k$$

introduced in [13]. Here S_0 is chosen as the nominal value of the process to be monitored. Here $S_0 = \mathbb{E}[r_k] = 0$, since the residual is expected to be 0 under

nominal behavior. The performance of the EWMA is said to be similar to the one of the CUSUM [14, p.419], which is why we choose to include EWMA in our comparison. The parameter β is the forgetting factor of the EWMA and it influences how much influence the new residual r_k has on the metric S_{k+1}. For the forgetting factor $\beta \in [0, 1]$ applies and typical values are $\beta \in [0.05, 0.25]$[14, p.423].
Usually the residual r_k has more than one element and in this case one has to use the multivariate exponentially weighted moving average (MEWMA) filter [15]

$$\begin{aligned} S_{k+1,\mathrm{M}} &= \beta r_k + (1-\beta) S_{k,\mathrm{M}} \\ S_{k+1} = S_{k+1,\mathrm{M}}^T \Sigma_{S_{k+1,\mathrm{M}}}^{-1} S_{k+1,\mathrm{M}} &\leq J_{k,D} \text{ for no alarms,} \end{aligned}$$

where $\Sigma_{S_{k+1,\mathrm{M}}}$ is the covariance matrix of $S_{k+1,\mathrm{M}}$ and $J_{k,D}$ is a time varying threshold. A simplification of the MEWMA is used

$$\begin{aligned} S_{k+1,\mathrm{M}} &= \beta r_k + (1-\beta) S_{k,\mathrm{M}} \\ S_{k+1} = S_{k+1,\mathrm{M}}^T S_{k+1,\mathrm{M}} &\leq J_D \text{ for no alarms,} \end{aligned}$$

with a constant threshold and no scaling by the covariance matrix, because it is easier to compare with the other detectors. Since the MEWMA is linear, we can include it into the state

$$\begin{bmatrix} \mu_{k+1} \\ S_{k+1,\mathrm{M}} \end{bmatrix} = \begin{bmatrix} A_e & 0 \\ \beta C_e & (1-\beta) I \end{bmatrix} \begin{bmatrix} \mu_k \\ S_k \end{bmatrix} + \begin{bmatrix} B_e \\ \beta D_e \end{bmatrix} a_k \text{ with } \mu_0 = const \text{ and } S_{0,\mathrm{M}} = 0.$$

Now define $\mu_{k,s}^T = [\mu_k^T \ S_{k,\mathrm{M}}^T]$ to get

$$\begin{aligned} \mu_{k+1,s} &= A_{e,s} \mu_{k,s} + B_{e,s} a_k \\ r_{k,s} &= C_{e,s} \mu_{k,s} + D_{e,s} a_k \end{aligned}$$

with

$$\begin{aligned} A_{e,s} &= \begin{bmatrix} A_e & 0 \\ \beta C_e & (1-\beta) I \end{bmatrix}, B_{e,s} = \begin{bmatrix} B_e \\ \beta D_e \end{bmatrix} \\ C_{e,s} &= \begin{bmatrix} \beta C_e & (1-\beta) I \end{bmatrix}, D_{e,s} = \beta D_e. \end{aligned}$$

Note that this is equivalent to a new system with a stateless detector.
Therefore everything derived for the stateless detector in our work is also valid for the MEWMA, because we can create this system with a stateful detector "hidden" in the system states and use the stateless anomaly detector

$$S_{k+1} = ||r_{k,s}||_2^2.$$

Chapter 3

Methods

The main goal is to compare the three anomaly detectors and see which one limits the adversary the most. Therefore the worst-case impact of the attack on the system under the different anomaly detectors has to be determined. The attacker does not want to be detected and this can be formulated as an optimal control problem where we want to maximize the influence of the attack on the system under the constraint that the attack is not detected by the anomaly detector.
This chapter represents the main part of our work, since it provides the theoretical results to compare the detectors. It defines thresholds for the anomaly detectors and the optimal control problem for maximizing the attack impact while remaining stealthy, which is a non-convex optimization problem. An analytical solution for a special case and the steady state impact is presented as well as a relaxation, which results in a convex optimization problem. At the end of the chapter, two ways of determining a novel stealthy bang-bang attack are proposed.

3.1 Thresholds for the Anomaly Detectors

Before we can characterize the worst-case attacks one has to choose a threshold J_D for each anomaly detector. One way to choose the thresholds is proposed here.
The proposed threshold for the stateless anomaly detector is

$$J_D = J_{th} = \mathbb{E}[||r_k||_2^2] + i\sqrt{\mathrm{Var}[||r_k||_2^2]} \tag{3.1}$$

with $i \in \{3, 4, 5\}$ and according to Chebyshev's inequality 11.1111%, 6.25% or 4% of the nominal values of $||r_k||_2^2$ will lie outside this threshold. This should therefore prevent too many false alarms caused by noise and at the same time should not be too high to detect dubious distortions of the system for the stateless detector. Henceforth we call the threshold J_D of the stateless detector J_{th} as defined in (3.1).
In the following we present how to determine the expected value and variance of $||r_k||_2^2$. The assumptions and procedure of determining this threshold is closely related to parts of the Kalman filter derivation given in [16, p. 310ff].

Assume for the process and measurement noise $w_k \sim \mathcal{N}(0, \Sigma_{w_k})$ and $v_k \sim \mathcal{N}(0, \Sigma_{v_k})$. Moreover the process and measurement noise are uncorrelated and the initial state is also a stationary Gaussian random variable, which is uncorrelated to the noise.
From these assumptions we know that $r_k \sim \mathcal{N}(0, \Sigma_{r_k})$ with

$$\Sigma_{r_k} = \mathbb{E}[r_k r_k^T] = C^T \Sigma_{e_k} C + \Sigma_{v_k},$$

where $\Sigma_{e_k} = \mathbb{E}[e_k e_k^T] = \mathbb{E}[(x_k - z_k)(x_k - z_k)^T]$ is the error covariance matrix and

$$\Sigma_{e_{k+1}} = (A - LC)\Sigma_{e_k}(A - LC)^T + \Sigma_{w_k}.$$

If the Kalman gain L of the Kalman filter is used, Σ_{e_k} will be minimal and has the following recursion, which is given by the discrete Riccati equation

$$\Sigma_{e_{k+1}} = A\Sigma_{e_k}A^T - A\Sigma_{e_k}C^T(C\Sigma_{e_k}C^T + \Sigma_{v_k})^{-1}C\Sigma_{e_k}A^T + \Sigma_{w_k}.$$

To get a constant threshold we further assume that w_k and v_k are stationary processes ($\Sigma_{w_k} = \Sigma_w$ and $\Sigma_{v_k} = \Sigma_v$), so Σ_{e_k} will converge to

$$\Sigma_e = A\Sigma_e A^T - A\Sigma_e C^T(C\Sigma_e C^T + \Sigma_v)^{-1}C\Sigma_e A^T + \Sigma_w.$$

This then leads to $r_k \sim \mathcal{N}(0, \Sigma_r)$ with

$$\Sigma_r = \mathbb{E}[r_k r_k^T] = C^T \Sigma_e C + \Sigma_v.$$

The signal considered in the stateless anomaly detector is $||r_k||_2^2$ and with [17] we get

$$\begin{aligned}\mathbb{E}[||r_k||_2^2] &= tr(C^T\Sigma_e C + \Sigma_v) = tr(\Sigma_r)\\ \mathrm{Var}[||r_k||_2^2] &= 2tr(\Sigma_r\Sigma_r)\end{aligned}$$

as the expected value and the variance of $||r_k||_2^2$ under no attack to determine the threshold J_{th}, because $r_k \sim \mathcal{N}(0, \Sigma_r)$.
It is more difficult to obtain thresholds for the CUSUM and MEWMA detector, but we can use the threshold of the stateless detector J_{th} as a starting point. The easiest way is to choose $J_D = J_{th}$ also as a threshold for CUSUM and MEWMA. Another approach is to consider that if the stateless detector triggers an alarm for $||r_k||_2^2 > J_{th}$ the stateful detectors should do the same. Considering the worst-case scenario under a stateless detector ($||r_k||_2^2 = J_{th}$) with the CUSUM and MEWMA detector, which have no information about the past ($S_k = 0$) yet, we get for the CUSUM detector

$$\begin{aligned}S_{k+1} = \max(0, ||r_k||_2^2 - \delta) &= J_{th} - \delta\\ \Rightarrow J_{D,\mathrm{CUSUM}} &= J_{th} - \delta\end{aligned}$$

and similarly for the MEWMA detector

$$\begin{aligned}S_{k+1}^T S_{k+1} = \beta^2||r_k||_2^2 &= \beta^2 J_{th}\\ \Rightarrow J_{D,\mathrm{MEWMA}} &= \beta^2 J_{th}\end{aligned}$$

to directly detect $||r_k||_2^2 > J_{th}$.
Therefore the later more investigated thresholds for CUSUM are chosen to be

- $J_D = J_{th}$
- $J_D = J_{th} - \delta$

and for MEWMA

- $J_D = J_{th}$
- $J_D = \beta^2 J_{th}$

Note that we refer to J_D as the threshold determined for the squared norm of a residual. If we want to consider the norm of a residual the threshold $\sqrt{J_D}$ is used.
Additionally we obtain a lower bound for the CUSUM forgetting factor δ, since we know

$$\begin{aligned} &\mathbb{E}[||r_k||_2^2 - \delta] < 0 \\ \Leftrightarrow\ &\mathbb{E}[||r_k||_2^2] < \delta \end{aligned}$$

to avoid too many false alarms. In the following we choose

$$\delta = \mathbb{E}[||r_k||_2^2] + \sqrt{\mathrm{Var}[||r_k||_2^2]}, \tag{3.2}$$

as an intuitive choice for the forgetting factor to eliminate the noise in the residual and to avoid false alarms.

3.2 The Optimization Problems Being Solved by Adversaries

The adversary is considered to have one of two objectives during his attack. The first objective is to maximize the attack impact on the whole trajectory during the attack. This can be expressed as the optimization problem

$$\max_{\{a_k\}_{k=0}^{N-1}} \sum_{k=0}^{N-1} ||x_{k+1}||_2^2 \text{ s.t. } S_{k+1} \leq J_D \ \forall k \in \{0, \ldots, N-1\}. \tag{3.3}$$

The second objective is to maximize the impact on the final state of the system

$$\max_{\{a_k\}_{k=0}^{N-1}} ||x_N||_2^2 \text{ s.t. } S_{k+1} \leq J_D \ \forall k \in \{0, \ldots, N-1\}. \tag{3.4}$$

Both optimization problems have the constraint that the attack should not trigger any alarms in the anomaly detectors.
When designing the attack we neglect the process and measurement noise w_k and v_k, although we used it to determine the thresholds for the detectors. The reason for this is that we want to determine the worst-case impact on the system, which would be diminished by the addition of noise.
The optimization problem can be transformed into another form by expressing the extended states μ_k and residuals r_k by using only μ_0 and the attacks

$a_0, a_1, \ldots, a_{N-1}$

$$\mu_k = A_e^k \mu_0 + \sum_{i=0}^{k-1} A_e^{k-1-i} B_e a_i$$
$$r_k = C_e A_e^k \mu_0 + \sum_{i=0}^{k-1} C_e A_e^{k-1-i} B_e a_i + D_e a_k.$$

By stacking the extended states into $\mu = [\mu_1^T, \ldots, \mu_N^T]^T \in \mathbb{R}^{2Nn}$, the residuals into $r = [r_0^T, r_1^T, \ldots, r_{N-1}^T]^T \in \mathbb{R}^{Np}$ and the attack into $a = [a_0^T, a_1^T, \ldots, a_{N-1}^T]^T \in \mathbb{R}^{Nm}$, where we consider $a_k = 0$ for $k \geq N$, we get

$$\mu = A_c \mu_0 + B_c a$$

with

$$A_c = \begin{bmatrix} A_e \\ A_e^2 \\ \vdots \\ A_e^N \end{bmatrix} \quad \text{and} \quad B_c = \begin{bmatrix} B_e & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ A_e B_e & B_e & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & & & \ddots & & \vdots \\ A_e^{N-1} B_e & A_e^{N-2} B_e & A_e^{N-3} B_e & \cdots & A_e B_e & B_e \end{bmatrix}$$

for the problem in (3.3) and

$$A_c = A_e^N \text{ and } \quad B_c = \begin{bmatrix} A_e^{N-1} B_e & A_e^{N-2} B_e & A_e^{N-3} B_e & \cdots & A_e B_e & B_e \end{bmatrix}$$

for optimization problem (3.4), where we are only interested in the final state $\mu = \mu_N$ and

$$r = C_c \mu_0 + D_c a \tag{3.5}$$

with

$$C_c = \begin{bmatrix} C_e \\ C_e A_e \\ \vdots \\ C_e A_e^{N-2} \\ C_e A_e^{N-1} \end{bmatrix} \text{ and } D_c = \begin{bmatrix} D_e & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ C_e B_e & D_e & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & & \ddots & & \vdots \\ C_e A_e^{N-3} B_e & C_e A_e^{N-4} B_e & \cdots & D_e & \mathbf{0} \\ C_e A_e^{N-2} B_e & C_e A_e^{N-3} B_e & \cdots & C_e B_e & D_e \end{bmatrix} \tag{3.6}$$

for both optimization problems, where $\mathbf{0}$ are zero matrices of appropriate dimensions.
Now we have equations for μ and r, but we want to maximize the states x_k so we define $x = [x_1^T, x_2^T, \ldots, x_N^T]^T$ and

$$x = T_x \mu$$

with

$$T_x = \begin{bmatrix} [I_n \ \mathbf{0}_n] & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & [I_n \ \mathbf{0}_n] & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & [I_n \ \mathbf{0}_n] \end{bmatrix}$$

for problem (3.3) and

$$T_x = \begin{bmatrix} I_n & \mathbf{0}_n \end{bmatrix}$$

for the second optimization problem (3.4). Here I_n and $\mathbf{0}_n$ are the identity and zero matrix of dimension n, respectively.
This leads to an alternative formulation of the optimization problems, which is shown here for problem (3.3)

$$\begin{aligned}\max_a \sum_{k=0}^{N-1} ||x_{k+1}||_2^2 &= \max_a x^T x = \max_a \mu^T T_x^T T_x \mu \\ &= \max_a \mu_0^T A_c^T T_x^T T_x A_c \mu_0 + 2\mu_0^T A_c^T T_x^T T_x B_c a + a^T B_c^T T_x^T T_x B_c a \\ &= \max_a a^T B_c^T T_x^T T_x B_c a + 2\mu_0^T A_c^T T_x^T T_x B_c a\end{aligned}$$

subject to $S_{k+1} \leq J_D \ \forall k \in \{0, \ldots, N-1\}$. Problem (3.4) has exactly the same form, but the matrices A_c, B_c and T_x have to be chosen accordingly.
This form indicates the non-convexity of the optimization problem. Clearly $B_c^T T_x^T T_x B_c$ is a positive semi-definite matrix. This shows we maximize a convex function over convex constraints[1], which leads to a non-convex optimization problem. Therefore the optimal solution can not be obtained easily using standard solvers, because non-convex problems do not have the characteristic that every local optimum is a global optimum as well.

3.2.1 Boundedness of the Problems

The boundedness of the problems is investigated in this section. Intuitively the problems are unbounded, if the adversary can create an attack that influences x, but simultaneously leads to $r = 0$. This leads to the necessary and sufficient condition

$$\ker(D_c) \subseteq \ker(B_c)$$

for the problems to be bounded (see Lemma 9 in [2] to get the proof idea). By investigating the structure of B_c and D_c one can see the last column of both matrices contains only zero matrices and B_e and D_e, respectively. So we get the necessary condition

$$\ker(D_e) \subseteq \ker(B_e)$$

otherwise one could create an attack as $a_k = 0 \ \forall k < N-1$ and $a_{N-1} \in \ker(D_e)$ to drive the system to an unbounded state.
Recall the special form of B_a and D_a mentioned in (2.1). This form leads in many cases to the fact that the attacker can create an attack, which has no influence on the residual, but results in an unbounded system state, i.e. $\ker(D_e) \nsubseteq \ker(B_e)$. Therefore we almost always have an unbounded problem for an attacker that attacks only for a limited amount of time, here N steps. What does this mean for the adversary?
Theoretically the adversary can drive the plant's state to infinity in a limited

[1] The three anomaly detectors lead to convex constraints as shown in Section 3.2.2.

amount of time without being detected in this time, but if he does the attack will definitely be detected one time step after it ended. The reason for that is that the attack will lead to an unbounded state x_N so that the next residual r_N will also grow without bound, which triggers an alarm in the anomaly detector. i.e.

$$||x_N||_2^2 \to \infty \Rightarrow ||r_N||_2^2 = ||C_e\mu_N||_2^2 \to \infty \Rightarrow S_{N+1} \to \infty.$$

Considering now that it is practically impossible to drive the system to an unbounded state in a finite time and that the attacker also wants to be stealthy after the attack, to for example attack again, the adversary has to take the aftermath of his attack into account to stay undetected.
Therefore we can define the new problems

$$\max_a \sum_{k=0}^{N-1} ||x_{k+1}||_2^2 \text{ s.t. } S_{k+1} \leq J_D \ \forall k \in \{0, \dots, N + \Delta N - 1\}$$

and

$$\max_a ||x_N||_2^2 \text{ s.t. } S_{k+1} \leq J_D \ \forall k \in \{0, \dots, N + \Delta N - 1\}$$

where the adversary considers the aftermath of his attack until the point $N + \Delta N - 1$. In this case we redefine the residual vector $r = [r_0^T \ \cdots, r_N, \ \cdots, r_{N+\Delta N-1}]^T$ and the equations (3.5) and (3.6) to

$$r = C_c\mu_0 + D_c a$$

where C_c and D_c are changed to

$$C_c = \begin{bmatrix} C_e \\ C_eA_e \\ \vdots \\ C_eA_e^{N-1} \\ C_eA_e^N \\ \vdots \\ C_eA_e^{N+\Delta N-1} \end{bmatrix}$$

and

$$D_c = \begin{bmatrix} D_e & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ C_eB_e & D_e & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & & \ddots & & \vdots \\ C_eA_e^{N-2}B_e & C_eA_e^{N-3}B_e & \cdots & C_eB_e & D_e \\ C_eA_e^{N-1}B_e & C_eA_e^{N-2}B_e & \cdots & C_eA_eB_e & C_eB_e \\ \vdots & & \ddots & & \vdots \\ C_eA_e^{N+\Delta N-2}B_e & C_eA_e^{N+\Delta N-3}B_e & \cdots & C_eA_e^{\Delta N}B_e & C_eA_e^{\Delta N-1}B_e \end{bmatrix}.$$

Recall that $a_k = 0 \ \forall k \geq N$. Choosing ΔN appropriately will result in a stealthy attack, that is also not detected for any $k \in \mathbb{N}_0$. From now on we assume that the optimization problems are bounded, i.e. $\ker(D_c) \subseteq \ker(B_c)$.

Now that the problems are defined, we will propose solutions to them, but due to the non-convexity and the N or $N+\Delta N$ constraints it is not trivial to obtain a closed form solution.
In the next two sections we give a closed-form solution for a special case of the CUSUM detector and the impact on the steady state with different anomaly detectors. After that a relaxation and convex reformulation are presented and finally a stealthy bang-bang attack is introduced.

3.2.2 Convexity of the Detectors

In this section, we want to show that the three investigated detectors actually represent convex constraints for the optimization problem. Starting with the residual recall

$$\begin{aligned} r_k &= C_e A_e^k \mu_0 + \sum_{i=0}^{k-1} C_e A_e^{k-1-i} B_e a_i + D_e a_k \\ &= C_{c_k} \mu_0 + D_{c_k} a \end{aligned}$$

with

$$C_{c_k} = C_e A_e^k \text{ and } D_{c_k} = \begin{bmatrix} C_e A_e^{k-1} Be & C_e A_e^{k-2} Be & \dots & C_e Be & De & 0 & \dots & 0 \end{bmatrix}.$$

The squared Euclidean norm of the residual is then given by

$$||r_k||_2^2 = \mu_0^T C_{c_k}^T C_{c_k} \mu_0 + a^T D_{c_k}^T D_{c_k} a + 2\mu_0 C_{c_k}^T D_{c_k} a.$$

which are convex quadratic inequality constraints

$$S_{k+1} = \frac{1}{2} a^T Q_k a + b_k^T a + c_k \leq J_{th} \ \forall k$$

with

$$Q_k = 2D_{c_k}^T D_{c_k},\ b_k^T = 2\mu_0 C_{c_k}^T D_{c_k} \text{ and } c_k = \mu_0^T C_{c_k}^T C_{c_k} \mu_0,$$

since Q_k is a positive semi-definite matrix. Therefore a stateless detector has convex constraints and it automatically follows that the MEWMA detector also results in convex constraints, since the MEWMA detector can be reformulated into a stateless detector as shown in Section 2.3.
The same is true for the CUSUM detector, which is proven by induction in the following. We want to prove that S_k represents convex constraints on a for the CUSUM.
First we prove that S_0 is convex, since $S_0 = 0$. Here S_0 is constant and therefore convex and concave in a at the same time, which shows S_0 is convex. Now assume S_k is convex and let us prove that S_{k+1} is convex as well.
We know that $||r_k||_2^2$ is convex and furthermore $-\delta$ is convex because it is constant. Using [18, p.79] we get that the nonnegative weighted sum of convex functions is convex and taking the maximum of two convex functions also results in a convex function. Hence, $S_k + ||r_k||_2^2 - \delta$ is convex and because of that $S_{k+1} = \max(0, S_k + ||r_k||_2^2 - \delta)$ is also convex, which concludes the proof by induction that S_k represents convex constraints for all k.
Hence, the three detectors result in convex constraints on a.

3.3 Analytical Solution for CUSUM

The analytical solution is based on the fact that we have to consider only one constraint for the optimization if $\delta = 0$.
If we set the CUSUM forgetting factor δ to zero we get

$$S_{k+1} = \max(0, S_k + ||r_k||_2^2) \leq J_D$$

with $S_0 = 0$. Obviously $||r_k||_2^2 \geq 0 \, \forall k$ so we can neglect the max-operator

$$S_{k+1} = S_k + ||r_k||_2^2 = \sum_{k=0}^{k} ||r_k||_2^2 \leq J_D.$$

With this we can reduce the N or $N + \Delta N$ constraints to just one constraint. If $S_N = r^T r \leq J_D$ or $S_{N+\Delta N} = r^T r \leq J_D$ then we know all other constraints are fulfilled as well, because we sum nonnegative values up. The optimization problem is then given by

$$\begin{aligned} &\max_a a^T B_c^T T_x^T T_x B_c a + 2\mu_0^T A_c^T T_x^T T_x B_c a \\ &\text{s.t. } r^T r = a^T D_c^T D_c a + 2\mu_0^T C_c^T D_c a + \mu_0^T C_c^T C_c \mu_0 \leq J_D. \end{aligned}$$

Note that we do not make any assumption whether we solve the worst-case problem for the whole trajectory or just for the final state, so this solution can be used for maximizing either the impact on the whole trajectory or the impact on the final state. Furthermore the aftermath of the attack can also be considered.
This is a maximization of a quadratic convex function over one quadratic convex constraint and necessary and sufficient conditions for a global optimum in this case are given in [19]. The problem is solved similar to the bias injection attack in [2], where these conditions are also used. To obtain an analytical closed form solution we have to assume $\mu_0 = 0$. For $\mu_0 = 0$ The problem results in

$$\max_a a^T B_c^T T_x^T T_x B_c a \text{ s.t. } r^T r = a^T D_c^T D_c a \leq J_D. \tag{3.7}$$

The conditions for a global optimum are

$$\begin{aligned} 0 &= (B_c^T T_x^T T_x B_c - \lambda D_c^T D_c) a^* \\ 0 &= a^{*T} D_c^T D_c a^* - J_D \\ 0 &\geq x^T (B_c^T T_x^T T_x B_c - \lambda D_c^T D_c) x \text{ for } x \neq 0. \end{aligned}$$

The first condition shows that λ has to be a generalized eigenvalue of the matrix pencil $(B_c^T T_x^T T_x B_c, \; D_c^T D_c)$ and the optimal attack has to be $a^* = \kappa v$, where v is the corresponding unit-norm eigenvector to λ and κ is a scaling factor.
From the third condition we get that λ has to be the largest generalized eigenvalue of the matrix pencil $(B_c^T T_x^T T_x B_c, \; D_c^T D_c)$ according to Lemma 10 of [2]. The second conditions gives κ as

$$\kappa = \pm \sqrt{\frac{J_D}{v^T D_c^T D_c v}}.$$

The sign used here is arbitrary, because κ appears quadratically in the second condition.
The optimal attack a^* is therefore given by

$$a^* = \pm\sqrt{\frac{J_D}{v^T D_c^T D_c v}}v$$

where v is the corresponding unit-norm eigenvector to the largest generalized eigenvalue λ of the matrix pencil $(B_c^T T_x^T T_x B_c,\ D_c^T D_c)$ and the optimal objective value is given as ${a^*}^T B_c^T T_x^T T_x B_c a^* = \lambda J_D$.
Note that the non-convex optimization problem for the CUSUM with $\delta = 0$ and $\mu_0 \neq 0$ can be reformulated as a semidefinite program to obtain a numerically efficient solution [18, p.653f].

3.4 Steady State Influence of the Attack

Since it is difficult to solve the problems (3.3) and (3.4) in general, we are investigating the steady state behavior of each detector and how much impact the attack can have on the steady state while remaining stealthy. The steady state can be seen as maximizing the final state of the attack, where $N \to \infty$ with $\Delta N = 0$. This is closely related to the bias injection attack presented in [2], where the influence of a false data-injection attack on the steady state is maximized.
Firstly we define the steady state for the extended state and residual.

$$\begin{aligned} x_\infty &= T_x(I - A_e)^{-1}B_e a_\infty =: G_{xa}a_\infty \\ r_\infty &= (C_e(I - A_e)^{-1}B_e + D_e)a_\infty =: G_{ra}a_\infty. \end{aligned}$$

The detectors are in steady state if $S_{k+1} = S_k$.

- **Stateless Detector:**
  Clearly the steady state is given by

  $$S_\infty = ||r_\infty||_2^2 \leq J_{th}.$$

- **MEWMA:**
  For the MEWMA steady state we get

  $$\begin{aligned} S_{k+1} &= \beta r_k + (1-\beta)S_k = S_k \\ \Leftrightarrow S_\infty &= r_\infty \\ \Rightarrow ||S_\infty||_2^2 &= ||r_\infty||_2^2 \leq J_D. \end{aligned}$$

  Here we get the same steady state condition as in the stateless detector case independent of the β, if $J_D = J_{th}$ is chosen. But one can also choose $J_D = \beta^2 J_{th}$ to get a dependence on the forgetting factor.

- **CUSUM:**
  For the CUSUM we have to consider two cases, when we want to examine the steady state

1. $S_\infty = 0$
   With this condition we get

$$S_\infty = \max(0, S_\infty + ||r_\infty||_2^2 - \delta) \Rightarrow ||r_\infty||_2^2 \leq \delta \text{ for } S_\infty = const.$$

2. $S_\infty > 0$
   With this condition we get

$$S_\infty = \max(0, S_\infty + ||r_\infty||_2^2 - \delta) \Rightarrow ||r_\infty||_2^2 = \delta \text{ for } S_\infty = const.$$

It follows that either $||r_\infty||_2^2 \leq \delta$ or $||r_\infty||_2^2 = \delta$ for the attack to be undetected under a CUSUM detector in the steady state. Note that the threshold does not depend on the actual threshold J_D. Similar results are shown and used to design an attack in [1], when the detector is in steady state.

All detectors lead to conditions on the steady state residual, so that we can use the optimization problem

$$\begin{aligned} &\max_{a_\infty} ||G_{xa} a_\infty||_2^2 \text{ s.t. } ||G_r a_\infty||_2^2 \leq J \\ \Leftrightarrow &\max_{a_\infty} a_\infty^T G_{xa}^T G_{xa} a_\infty \text{ s.t. } a_\infty^T G_{ra}^T G_{ra} a_\infty \leq J \end{aligned}$$

where J is the constraint on the steady state residual by a given detector, to see how much influence an attack can have on the steady state with different detectors while remaining stealthy. Note that this optimization problem is bounded if and only if $\ker(G_{ra}) \subseteq \ker(G_{xa})$ (see Lemma 9 in [2]) and has exactly the same form as (3.7), so the solution to the steady state problem can be obtained in the same way.
The objective value or impact on the system is then $||G_{xa} a_\infty^*||_2^2 = \lambda J$, where λ is the largest generalized eigenvalue of the matrix pencil $(G_{xa}^T G_{xa},\ G_{ra}^T G_{ra})$. So one can see that if MEWMA uses the same threshold as the stateless detector, it is not more sensitive to attacks than the stateless detector in the steady state case, while CUSUM does not depend on the threshold chosen in steady state but on the forgetting factor. Therefore CUSUM reduces the impact on the steady state more than the stateless and MEWMA detector, if all have the same constant threshold.
The steady state influence for different thresholds and forgetting factors is further investigated in Chapter 4, where we have actual models at hand and can calculate the steady state impact.

3.5 Relaxation and Convex Reformulation

Due to the non-convexity and many constraints it is difficult to solve the optimal control problems both analytically and numerically. In case of a numerical solution it is hard to know if the solution found is the global optimum for a non-convex problem. Therefore a relaxation is presented in this section that results in a convex optimization problem.
In (3.3) and (3.4) we are maximizing a convex objective function over convex constraints and for a convex optimization we have to maximize a concave function over convex constraints. Considering now that the squared Euclidean norm

results in a convex objective function, we are trying to find an alternative for the Euclidean norm, that gives also upper and lower bounds on the Euclidean norm and simultaneously yields a concave objective function. A natural choice is the infinity norm for upper and lower bounds, since

$$||x||_\infty \leq ||x||_2 \leq \sqrt{n}||x||_\infty \text{ for } x \in \mathbb{R}^n.$$

As shown below the infinity norm can also be used to formulate convex optimization problems with a concave objective function, which solve the relaxed problem.
We get the relaxed problems

$$\max_a ||x||_\infty \text{ s.t. } S_{i+1} \leq J_D \ \forall i \in \{0, \dots, N + \Delta N - 1\}$$

and

$$\max_a ||x_N||_\infty \text{ s.t. } S_{i+1} \leq J_D \ \forall i \in \{0, \dots, N + \Delta N - 1\}$$

where $\Delta N \geq 0$.
Now using that each state $x_k = [x_{1,k}, \cdots, x_{n,k}]^T$ and
$||x||_\infty = \max(|x_{1,1}|, \cdots, |x_{n,N}|)$ we get

$$\max_a ||x||_\infty = \max_{\substack{l \in \{1,\cdots,n\}, \\ k \in \{1,\cdots,N\}}} \max_a |x_{l,k}| \text{s.t. } S_{i+1} \leq J_D$$
$$\forall i \in \{0, \cdots, N + \Delta N - 1\}$$

and

$$\max_a ||x_N||_\infty = \max_{l \in \{1,\cdots,n\}} \max_a |x_{l,N}| \text{s.t. } S_{i+1} \leq J_D$$
$$\forall i \in \{0, \dots, N + \Delta N - 1\}$$

The problems are solved element-wise, where for example we obtain $x_{l,k}$ by

$$x_{l,k} = e_i^T x = e_i^T T_x (A_c \mu_0 + B_c a).$$

Here e_i^T is the ith row of an identity matrix of dimension of x, where i is chosen so that we get $x_{l,k}$ out of x. Furthermore we are splitting the problems further up, to eliminate the non-smooth absolute value function in the objective function

$$\max_a ||x||_\infty = \max_{\substack{l \in \{1,\cdots,n\}, \\ k \in \{1,\cdots,N\}}} \max_{j \in \{1,2\}} \max_a (-1)^j x_{l,k} \text{ s.t. } S_{i+1} \leq J_D \tag{3.8}$$

$\forall i \in \{0, \cdots, N + \Delta N - 1\}$ and

$$\max_a ||x_N||_\infty = \max_{l \in \{1,\cdots,n\}} \max_{j \in \{1,2\}} \max_a (-1)^j x_{l,N} \text{ s.t. } S_{i+1} \leq J_D \tag{3.9}$$

$\forall i \in \{0, \dots, N + \Delta N - 1\}$.
In these optimization problems we have a linear objective function in the attack vector a

$$(-1)^j x_{l,k} = (-1)^j e_i^T T_x (A_c \mu_0 + B_c a).$$

Linear functions are simultaneously convex and concave, hence we obtain a convex optimization problem, where a local optimum is a global optimum as well.
This has the advantage that the solution found numerically is also the global solution, but the disadvantage here is that one has to solve $2Nn$ or $2n$ convex optimization problems instead of one non-convex optimization problem, which does not scale well, if the attack horizon N is large.

3.5.1 Scaling of the Optimization for $\Delta N = 0$

As mentioned before the problem of maximizing the attack impact on the whole trajectory scales badly in the time horizon N, but in the case where the attacker does not consider the aftermath of his attack ($\Delta N = 0$) we do not need to solve the problem over the whole horizon due to the time invariance of the linear system.
Consider the two problems

$$\max_a ||x||_\infty \text{ s.t. } S_{k+1} \leq J_D \ \forall k \in \{0, \cdots, N-1\} \tag{3.10}$$

and

$$\max_a ||x_N||_\infty \text{ s.t. } S_{k+1} \leq J_D \ \forall k \in \{0, \cdots, N-1\} \tag{3.11}$$

with the optimal values F^* and F^*_{Final} and optimal solutions a^* and a^*_{Final}, respectively. Clearly both attacks fulfill the constraints $S_{k+1} \leq J_D$, hence they are feasible solutions for both problems. Furthermore we get that

$$F^* \geq F^*_{\text{Final}}$$

because x_N is contained in x.
Now consider we found a solution $F^* = ||x^*_k||_\infty \geq F^*_{\text{Final}}$ with $k < N$ and an attack a^* for (3.10). Due to the linearity and time invariance, we can design an attack for (3.11), so that $a_k = 0$ for $k \leq N-k-1$ and $a_k = a^*_{0:k-1}$ for $k > N-k-1$, where $a^*_{0:k-1}$ corresponds to the first k attacks in a^*. This attack would lead to $||x_N||_\infty = ||x^*_k||_\infty$ and also fulfill the constraints. Therefore we can always design an attack for (3.11), that results in $F^*_{\text{Final}} = F^*$, but only if $\Delta N = 0$ because the time shifted attack will not automatically guarantee that $S_{k+1} \leq J_D$ for $k \geq N$ since $a_k = 0$ for $k \geq N$.
This means that instead of solving $2Nn$ problems one can solve only $2n$ problems to obtain the optimal attack for (3.10) in the case of $\Delta N = 0$, which eliminates the scaling in N.

3.5.2 CUSUM Reformulation

For better numerical treatment, the CUSUM detector is reformulated, since the constraints contain a non-smooth function with the max-operator.
The optimization problem using the CUSUM detector can be formulated in a more general way as

$$\max_a f(a) \text{ s.t. } S_{k+1} = \max(0, S_k + ||r_k||_2^2 - \delta) \leq J_D, \ S_0 = 0 \tag{3.12}$$

and the reformulation used for the numerical solution is

$$\max_{a,\{P_k\}_{k=0}^{N+\Delta N-1}} f(a) \text{ s.t. } \begin{cases} P_{k+1} \geq 0 \\ P_{k+1} \geq P_k + ||r_k||_2^2 - \delta \\ P_{k+1} \leq J_D \\ P_0 = 0 \end{cases} \tag{3.13}$$

The fact that we get the same solution a for both formulations is proven in the following.
First of all one can easily see that $S_k \leq P_k \ \forall k$, if a is fixed and feasible in both (3.12) and (3.13). Assume now one obtained an optimal solution a^*_{CUSUM} for (3.12). This solution also fulfills the constraints of (3.13), since $P_k = S_k \leq J_D$ fulfills the constraints, which makes a^*_{CUSUM} a feasible solution for (3.13). But by solving (3.13) we might find a better solution, so that

$$f(a_r^*) \geq f(a^*_{\text{CUSUM}}). \tag{3.14}$$

Assume now one has an optimal solution $(a_k^*, \{P_k^*\}_{k=0}^{N+\Delta N-1})$ for (3.13) and one uses $S_k = P_k^*$ in (3.12). The P_k^*s might not fulfill the constraints of (3.12), but we can define a feasible sequence S_k from $\{P_k^*\}_{k=0}^{N+\Delta N-1}$ for (3.12) by using the lower bounds $P_{k+1}^* = \max(0, P_k^* + ||r_k||_2^2 - \delta)$, that does not change the value of the objective function of (3.13). Then $(a_r^*, \{P_k^*\}_{k=0}^{N+\Delta N-1})$ fulfills the constraints of (3.12) and is therefore a feasible solution for (3.12). But again we might find a better solution by solving (3.12) directly, so that

$$f(a^*_{\text{CUSUM}}) \geq f(a_r^*). \tag{3.15}$$

Therefore the inequalities (3.14) and (3.15) imply

$$f(a^*_{\text{CUSUM}}) = f(a_r^*),$$

which makes the problems equivalent and the reformulation valid.

3.6 Stealthy Bang-Bang Attacks

In this section we present a stealthy bang-bang attack to maximize the impact on the final state. The idea is to estimate an upper bound for the attacks $||a_k||_2 \leq c_a$ so that $S_{k+1} \leq J_D \ \forall k$. That leads to the new optimization problem

$$\begin{aligned} &\max_{\{a_k\}_{k=0}^{N-1}} ||x_N||_\infty \text{ s.t. } ||a_k||_2^2 \leq c_a^2 \ \forall k \in \{0, \cdots, N-1\} \\ \Leftrightarrow &\max_{\{a_k\}_{k=0}^{N-1}} ||x_N||_\infty = \max_{l \in \{1,\cdots,n\}} \max_{j \in \{1,2\}} \max_{\{a_k\}_{k=0}^{N-1}} (-1)^j x_{l,N} \\ &\text{s.t. } ||a_k||_2^2 \leq c_a^2 \ \forall k \in \{0, \cdots, N-1\}. \end{aligned}$$

Again we have to solve $2n$ problems to obtain the optimal attack.

3.6.1 Estimating an Upper Bound on the Attack

Estimating an upper bound c_a for the stateless and MEWMA detector is straight forward using the triangle inequality

$$\begin{aligned}
||r_k||_2 &= ||D_e a_k + \sum_{i=0}^{k-1} C_e A_e^{k-1-i} B_e a_i||_2 \\
&\leq ||D_e a_k||_2 + \sum_{i=0}^{k-1} ||C_e A_e^{k-1-i} B_e a_i||_2 \\
&\leq ||D_e||_2 ||a_k||_2 + \sum_{i=0}^{k-1} ||C_e A_e^{k-1-i} B_e||_2 ||a_i||_2 \\
&\leq c_a(||D_e||_2 + \sum_{i=0}^{k-1} ||C_e A_e^{k-1-i} B_e||_2) \leq \sqrt{J_D}.
\end{aligned}$$

Since this has to be fulfilled for all r_k with $0 \leq k \leq N-1$ we get

$$c_a \leq \frac{\sqrt{J_D}}{||D_e||_2 + \sum_{i=0}^{N-2} ||C_e A_e^{N-2-i} B_e||_2}, \tag{3.16}$$

where the spectral norm for matrices is used, which corresponds to the maximum singular value of the matrix. For the CUSUM detector it is not straight forward and a method to estimate an upper bound for the attacks in the CUSUM case is proposed in the following.
We consider two cases, where both assume $S_k + ||r_k||_2^2 - \delta \geq 0$
The first case assumes $S_k = 0$, which leads to

$$\begin{aligned}
&||r_k||_2^2 - \delta \leq J_D \\
&\Leftrightarrow ||r_k||_2^2 \leq J_D + \delta.
\end{aligned}$$

The second case assumes $S_k = J_D$

$$\begin{aligned}
J_D + ||r_k||_2^2 - \delta &\leq J_D \\
\Leftrightarrow ||r_k||_2^2 &\leq \delta.
\end{aligned}$$

It seems to be reasonable to assume that depending on S_k

$$||r_k||_2^2 \in [\delta, J_D + \delta]$$

is allowed for an attacker not to be detected. These values can then be used as upper bounds on $||r_k||_2^2$ to estimate an upper bound for the attack, but note that choosing $J_D + \delta$ might not result in a stealthy attack. Of course $||r_k||_2^2 < \delta$ is also allowed, but this will have a negligible impact on the system.
One can furthermore think of an attack which brings the residual to a constant value and stays undetected for the duration of the attack.

$$\begin{aligned}
&||r_k||_2^2 = \epsilon = const \in [\delta, J_D + \delta] \ \forall k \in [0, N] \\
&\Rightarrow \epsilon = \frac{J_D}{N} + \delta.
\end{aligned}$$

It is obvious that $\epsilon \to \delta$ if $N \to \infty$.
With these upper bounds on $||r_k||_2^2$ for CUSUM one can estimate an upper bound on the attack c_a with the methods for a stateless detector.
Ideally the problem with the upper bound would result in same solution as the attack on the final state in (3.9), but c_a is a conservative bound (see Section 4.1.4). Therefore we will get a different solution compared to (3.9), which only approximates this problem.

3.6.2 Solution to the Bounded Attack Problem

Pontryagins Maximum Principle (PMP) [20] is used to solve each one of the $2n$ optimization problems and one of the $2n$ problems is derived as

$$\max_{\{a_k\}_{k=0}^{N-1}} (-1)^j x_{l,N} = \max_{\{a_k\}_{k=0}^{N-1}} (-1)^j e_l^T T_x \mu_N \text{ s.t. } \mu_{k+1} = A_e\mu_k + B_e a_k \text{ and } ||a_k||_2^2 \leq c_a^2$$

Applying PMP leads to

$$\begin{aligned}
H(k, \mu_k, a_k, \lambda_{k+1}) &= \lambda_{k+1}^T (A_e\mu_k + B_e a_k) \text{ (Hamiltonian)} \\
\mu_{k+1} &= A_e\mu_k + B_e a_k, \ \mu_0 \text{ given (State Equation)} \\
\lambda_k &= A_e^T \lambda_{k+1} \text{ (Adjoint Equation)} \\
\lambda_N &= \frac{\delta (-1)^j e_l^T T_x \mu_N}{\delta \mu_N} = (-1)^j T_x^T e_l \text{ (Boundary Condition)} \\
a_k^* &= arg \max_{||a_k||_2^2 \leq c_a^2} H(k, \mu_k, a_k, \lambda_{k+1}) \\
&= arg \max_{||a_k||_2^2 \leq c_a^2} \lambda_{k+1}^T (A_e\mu_k + B_e a_k) \\
&= arg \max_{||a_k||_2^2 \leq c_a^2} \lambda_{k+1}^T B_e a_k \text{ (Pointwise Maximization)}
\end{aligned}$$

for $0 \leq k \leq N-1$.
The optimal attack a_k^* at time step k is similar to a bang-bang controller and therefore called bang-bang attack, where $||a_k||_2^2 = c_a^2 \ \forall k \in \{0, \cdots, N-1\}$. It is derived as

$$a_k^* = \begin{cases} c_a \dfrac{B_e^T \lambda_{k+1}}{\sqrt{\lambda_{k+1}^T B_e B_e^T \lambda_{k+1}}}, & \text{if } B_e^T \lambda_{k+1} \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.17}$$

This is derived by using the Karush-Kuhn-Tucker (KKT) conditions[18, p.243] to solve the pointwise maximization. Assume the optimization problem for fixed v

$$\max_u v^T u \text{ s.t. } ||u||_2^2 \leq c_a^2.$$

Then we get the Lagrangian with the KKT conditions and the KKT multiplier $\gamma \geq 0$

$$\begin{aligned}
L(u, \gamma) &= v^T u - \gamma(u^T u - c_a^2) \\
\Rightarrow \frac{\delta L}{\delta u} &= v - 2\gamma u = 0 \\
\Rightarrow u &= \frac{v}{2\gamma}.
\end{aligned}$$

This leads to the dual problem

$$\max_{\gamma} \frac{v^T v}{2\gamma} \text{ s.t. } \frac{v^T v}{4\gamma^2} \leq c_a^2.$$

So the smallest $\gamma \geq 0$ solves the problem

$$\begin{aligned} \gamma^* &= \frac{\sqrt{v^T v}}{2c_a} \\ \Rightarrow u^* &= \frac{c_a v}{\sqrt{v^T v}}. \end{aligned}$$

Setting $v = B_e^T \lambda_{k+1}$ and $u = a_k$ finally leads to the bang-bang solution in (3.17).
Since $\lambda_N = (-1)^j T_x^T e_l$ is a known constant for each problem, we can calculate all λ_k with the adjoint equation and determine the optimal attack for each problem analytically.
This problem can also be solved numerically for the objective function $||x_N||_2^2$, where the boundary condition for the adjoint variables changes to
$\lambda_N = 2T_x^T T_x \mu_N$. Here μ_N is an unknown parameter, but the problem can be solved with the shooting method by guessing a λ_N and iterating over the state equations until $\lambda_N = 2Tx^T T_x \mu_N$.

3.6.3 Infinity Norm for the Attack

Not only the Euclidean norm but also the infinity norm can be used to determine an upper bound for the attack a_k and to create a stealthy bang-bang attack. This is an alternative way to determine a bang-bang attack, which is not further investigated in the course of our work due to time limitations.
The optimization problem is given by

$$\begin{aligned} &\max_{\{a_k\}_{k=0}^{N-1}} ||x_N||_\infty \text{ s.t. } ||a_k||_\infty \leq c_a \ \forall k \in \{0, \cdots, N-1\} \\ \Leftrightarrow &\max_{\{a_k\}_{k=0}^{N-1}} ||x_N||_\infty = \max_{l \in \{1,\cdots,n\}} \max_{j \in \{1,2\}} \max_{\{a_k\}_{k=0}^{N-1}} (-1)^j x_{l,N} \\ &\text{s.t. } ||a_k||_\infty \leq c_a \ \forall k \in \{0, \cdots, N-1\}. \end{aligned}$$

where the upper bound c_a is determined in a similar way as described before

$$c_a \leq \frac{\sqrt{J_D}}{||D_e||_\infty + \sum_{i=0}^{N-2} ||C_e A_e^{N-2-i} B_e||_\infty}.$$

Again this problem is solved by applying PMP, which leads to

$$\begin{aligned}
H(k, \mu_k, a_k, \lambda_{k+1}) &= \lambda_{k+1}^T (A_e \mu_k + B_e a_k) \text{ (Hamiltonian)} \\
\mu_{k+1} &= A_e \mu_k + B_e a_k, \ \mu_0 \text{ given (State Equation)} \\
\lambda_k &= A_e^T \lambda_{k+1} \text{ (Adjoint Equation)} \\
\lambda_N &= \frac{\delta (-1)^j e_l^T T_x \mu_N}{\delta \mu_N} = (-1)^j T_x^T e_l \text{ (Boundary Condition)} \\
a_k^* &= arg \max_{||a_k||_\infty \leq c_a} H(k, \mu_k, a_k, \lambda_{k+1}) \\
&= arg \max_{||a_k||_\infty \leq c_a} \lambda_{k+1}^T (A_e \mu_k + B_e a_k) \\
&= arg \max_{||a_k||_\infty \leq c_a} \lambda_{k+1}^T B_e a_k \text{ (Pointwise Maximization)}
\end{aligned}$$

for $0 \leq k \leq N-1$. The pointwise maximization can be solved easily, if one considers that the constraint with the infinity norm leads to the fact that each element of the attack vector has to be smaller or equal to c_a.
The optimal attack is then given by

$$a_k^* = c_a \text{sign}(B_e^T \lambda_{k+1}), \tag{3.18}$$

where $\text{sign}(x)$ returns a vector with the sign of each element in x.

3.7 Residual-based Bang-Bang Attacks

In the following section, a stealthy residual-based bang-bang attack is presented by changing the constraints and assuming D_c is surjective, which means that every point $r \in \mathbb{R}^{(N+\Delta N)p}$ can be attained by using the right attack a.
If we do not only use the infinity norm for the objective function but also for the constraint on the residuals, we get a closed form solution for the stateless and MEWMA detector in case D_c is surjective. The new optimization problem is then

$$\max_a ||x||_\infty \text{ s.t. } ||r||_\infty \leq \sqrt{J_D}.$$

Recall

$$\begin{aligned}
x &= T_x A_c \mu_0 + T_x B_c a \\
r &= C_c \mu_0 + D_c a.
\end{aligned}$$

If D_c is surjective we can express all possible residuals in $\mathbb{R}^{(N+\Delta N)p}$ with the attack vector

$$a = D_c^+ r - D_c^+ C_c \mu_0 + a_{\text{Null}},$$

where r is the desired residual vector, $D_c^+ = D_c^T (D_c D_c^T)^{-1}$ is the pseudoinverse of D_c and $a_{\text{Null}} \in \ker(D_c)$. This can be seen as a re-parametrization of the attack vector dependent on the residual.
Inserting the new attack vector in the optimization problem leads to

$$\max_{r, a_{\text{Null}}} ||(T_x A_c - T_x B_c D_c^+ C_c)\mu_0 + T_x B_c D_c^+ r + B_c a_{\text{Null}}||_\infty \text{ s.t. } ||r||_\infty \leq \sqrt{J_D}.$$

It is obvious that this problem is unbounded if $B_c a_{\text{Null}} \neq 0$, which leads again to the condition $\ker(D_c) \subseteq \ker(B_c)$ for the problem to be bounded. Let us assume that the problem is bounded, so we can set $a_{\text{Null}} = 0$ without loss of generality. Moreover note that we can reformulate the constraint as

$$\begin{aligned} &||r||_\infty \leq \sqrt{J_D} \\ \Leftrightarrow\ &||r_k||_\infty \leq \sqrt{J_D}\ \forall k \in \{0, \cdots, N + \Delta N - 1\} \\ \Leftrightarrow\ &|r_{l,k}| \leq \sqrt{J_D}\ \forall k \in \{0, \cdots, N + \Delta N - 1\},\ \forall l \in \{1, \cdots, p\} \end{aligned}$$

i.e. the absolute value of each element $r_{l,k}$ in the stacked residual vector r has to be smaller or equal to $\sqrt{J_D}$.
Dividing the infinity norm objective function into several objective functions then leads to

$$\begin{aligned} &\max_{i \in \{1,\cdots,Nn\}} \max_{j \in \{1,2\}} \max_{||r||_\infty \leq \sqrt{J_D}} (-1)^j e_i^T ((T_x A_c - T_x B_c D_c^+ C_c)\mu_0 + T_x B_c D_c^+ r) \\ \Leftrightarrow\ &\max_{i \in \{1,\cdots,Nn\}} \max_{j \in \{1,2\}} \max_{||r||_\infty \leq \sqrt{J_D}} (-1)^j e_i^T T_x B_c D_c^+ r, \end{aligned}$$

The optimal solution is easily obtained by

$$\begin{aligned} r^* &= \sqrt{J_D} \text{sign}((-1)^{j^*} e_{i^*}^T T_x B_c D_c^+)^T \\ a^* &= D_c^+ r^* - D_c^+ C_c \mu_0, \end{aligned}$$

where i^* and j^* are the indices that lead to the worst-case attack for $||x||_\infty$ and $\text{sign}(x)$ returns a vector with the sign of each element of x. The optimal objective value is

$$(-1)^{j^*} e_{i^*}^T (T_x A_c - T_x B_c D_c^+ C_c)\mu_0 + \sqrt{J_D} ||e_{i^*}^T T_x B_c D_c^+||_1.$$

Certainly the case of maximizing the final state $||x_N||_\infty$ can also be solved in the same way by choosing A_c, B_c and T_x accordingly.
This solution is only possible if we use the infinity norm for the residuals and have a surjective D_c matrix, therefore we give two necessary conditions for D_c to be surjective

- $(N + \Delta N)m \leq Np$ or $m \geq (1 + \lceil \frac{\Delta N}{N} \rceil)p$, where $\lceil x \rceil$ is the ceiling operator.
- D_e has to have full row rank.

The first condition is necessary for D_c to have full row rank, which is one condition for D_c to be surjective, while the second condition is necessary so that we can reach all possible $r_0 \in \mathbb{R}^p$.

3.7.1 D_e has full row rank

If only the condition that D_e has full row rank is fulfilled, we can characterize a bang-bang attack similar to the one in Section 3.6.3. If D_e has full row rank we can reach every $r_k \in \mathbb{R}^p\, \forall k \leq N - 1$ with a_k independent of μ_k, because the pseudoinverse $D_e^+ = D_e^T (D_e D_e^T)^{-1}$ exists so that the attack

$$a_k = D_e^+ r_k - D_e^+ C_e \mu_k + a_{\text{Null}} \tag{3.19}$$

leads to the desired r_k with $a_{\text{Null}} \in \ker(D_e)$. So all $r_k \in \mathbb{R}^p$ are reachable with the attack a_k and with (3.19) one can also define a new system with the residual as an input

$$\mu_{k+1} = A_e\mu_k + B_e a_k = (A_e - B_e D_e^+ C_e)\mu_k + B_e D_e^+ r_k + B_e a_{\text{Null}}$$

and use this equation to maximize the final state as it is done in Section 3.6.3. One of the $2n$ optimal control problems to solve is given by

$$\begin{aligned}
&\max_{\{r_k\}_{k=0}^{N-1}, a_{\text{Null}}} (-1)^j x_{l,N} = \max_{\{r_k\}_{k=0}^{N-1}, a_{\text{Null}}} (-1)^j e_l^T T_x \mu_N \\
&\text{s.t. } \mu_{k+1} = (A_e - B_e D_e^+ C_e)\mu_k + B_e D_e^+ r_k + B_e a_{\text{Null}} \text{ and } ||r_k||_\infty \leq \sqrt{J_D}
\end{aligned}$$

for $0 \leq k \leq N-1$. Before we present the optimal residuals for this problem note that we get here as well that $\ker(D_e) \subset \ker(B_e)$ for a bounded problem, because of the a_{Null} term in the state equation. We assume now that we have a bounded problem and $a_{\text{Null}} = 0$ without loss of generality.
The solution can again be found by applying PMP

$$\begin{aligned}
H(k, \mu_k, r_k, \lambda_{k+1}) &= \lambda_{k+1}^T((A_e - B_e D_e^+ C_e)\mu_k + B_e D_e^+ r_k) \text{ (Hamiltonian)} \\
\mu_{k+1} &= (A_e - B_e D_e^+ C_e)\mu_k + B_e D_e^+ r_k, \ \mu_0 \text{ given (State Equation)} \\
\lambda_k &= (A_e - B_e D_e^+ C_e)^T \lambda_{k+1} \text{ (Adjoint Equation)} \\
\lambda_N &= \frac{\delta(-1)^j e_l^T T_x \mu_N}{\delta \mu_N} = (-1)^j T_x^T e_l \text{ (Boundary Condition)} \\
r_k^* &= arg \max_{||r_k||_\infty \leq \sqrt{J_D}} H(k, \mu_k, r_k, \lambda_{k+1}) \\
&= arg \max_{||r_k||_\infty \leq \sqrt{J_D}} \lambda_{k+1}^T((A_e - B_e D_e^+ C_e)\mu_k + B_e D_e^+ r_k) \\
&= arg \max_{||r_k||_\infty \leq \sqrt{J_D}} \lambda_{k+1}^T B_e D_e^+ r_k \text{ (Pointwise Maximization)}
\end{aligned}$$

for $0 \leq k \leq N-1$. The optimal residual is then given by

$$r_k^* = \sqrt{J_D}\text{sign}((B_e D_e^+)^T \lambda_{k+1})$$

which is derived in the same way as a_k^* in (3.17). The optimal attack is obtained by entering r_k^* into equation (3.19). Note that this attack might not be stealthy, since it does not consider the aftermath of the attack, i.e. $\Delta N = 0$. Hence, the attack might be detected at time step $k = N$.

3.8 Summary

In this chapter the worst-case attacks are characterized as optimal control problems, where the attack signals a_k are interpreted as the systems input. The problems are non-convex so relaxations using the infinity norm for the objective function are proposed for better numerical treatment of the problems. Moreover a new attack, the bang-bang attack, is discovered.
All discussed attacks from this chapter are summarized in Table 3.1 with their respective objective function, constraints, consideration of the aftermath and some additional attack specific information.

Table 3.1: Summary of the Discussed Attacks

Attack Scheme	Objective Function	ΔN	Constraints	Additional Information
Analytical Solution to CUSUM	$\|\|x\|\|_2^2$ or $\|\|x_N\|\|_2^2$	≥ 0	$S_{N+\Delta N} \leq J_D$	$\delta = 0$
Steady State	$\|\|x_\infty\|\|_2^2$	$= 0$	$S_\infty \leq J_D$	$N \to \infty$
Relaxation	$\|\|x\|\|_\infty$ or $\|\|x_N\|\|_\infty$	≥ 0	$S_{k+1} \leq J_D$	$2Nn$ or $2n$ problems to solve
Bang-Bang Attack	$\|\|x_N\|\|_\infty$	$= 0$	$\|\|a_k\|\|_2 \leq c_a$ or $\|\|a_k\|\|_\infty \leq c_a$	Conservative problem
Bang-Bang Attack (Residual)	$\|\|x\|\|_\infty$ or $\|\|x_N\|\|_\infty$	≥ 0	$\|\|r\|\|_\infty \leq \sqrt{J_D}$	Not for CUSUM

In the further course of the thesis these worst-case attacks are used both in simulation and in an experiment to examine which detector reduces the attack impact in the best way, i.e. restrains the attacker the most in all cases considered.

Chapter 4

Simulations and Experiments

In this section the anomaly detectors are compared and evaluated. First a simple simulation example is used to perform a worst-case sensitivity analysis for each detector under attack. To see how good the theory applies to a realistic case, the anomaly detectors are then investigated in a realistic simulation and a real experiment with the quadruple tank process described in [21].
The attack impact under the stateless detector is used as a reference of the impact in both examples, because this detector is widely used in the control literature and has just one possible configuration, while we vary the forgetting factor and investigate two different thresholds for the CUSUM and MEWMA detectors. In this chapter we furthermore assume that the initial state is $\mu_0 = 0$, which is no restriction, because the models used are linearized models that should initially have no deviation from the system's state.

4.1 Simple Simulation Example

The example system is a slight modification of the system used in [7, p.28]. This example system is used to show the effect of the attack on the steady state, the whole trajectory, and the final state under different anomaly detectors. The effect of stealthy bang-bang attacks under different anomaly detectors is investigated as well. Therefore we do not consider the noise impact on the system since we want to determine the sensitivity of each detector under an attack, but the noise statistics are used to determine the thresholds for the detectors, since they have to be chosen to avoid too many false alarms according to the noise statistics (see Section 2.3).

4.1.1 Model Equations

This example is based on a nonlinear model, but since this should be just a simple example system, we will not consider the nonlinearities. This means we design the attack for the linearized system and also apply it to the linearized system in contrast to the quadruple tank process, where we design the attack for the linearized system and apply it to the nonlinear system.

The continuous-time nonlinear system is given by

$$\begin{aligned} \omega &= \dot{\theta}(t) \\ M\dot{\omega}(t) &= -D\omega(t) - b\sin(\theta(t)) + u(t), \end{aligned}$$

which describes generator dynamics as a normalized swing equation in the example from [7], where ω is the generator's frequency deviation and θ its phase-angle. After linearizing the system around the operating point $x = [\theta, \omega] = [0, 0]^T$ with $M = D = b = 1$ and sampling it with a sample time of $T_s = 1\,s$, the system is described by

$$\begin{aligned} x_{k+1} &= Ax_k + B\tilde{u}_k \\ y_k &= Cx_k \end{aligned}$$

with

$$A = \begin{bmatrix} 0.66 & 0.53 \\ -0.53 & 0.13 \end{bmatrix},\ B = \begin{bmatrix} 0.34 \\ 0.53 \end{bmatrix} \text{ and } C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

A state observer is used to control the system

$$\begin{aligned} z_{k+1} &= (A - LC)z_k + Bu_k + L\tilde{y}_k \\ u_k &= -Kz_k \\ r_k &= \tilde{y}_k - Cz_k \end{aligned},$$

where

$$K = \begin{bmatrix} 0.0556 & 0.3306 \end{bmatrix} \text{ and } L = \begin{bmatrix} 0.36 & 0.27 \\ -0.31 & 0.08 \end{bmatrix}$$

are obtained by designing a linear-quadratic Gaussian controller.
We assume now that the control input u_k and the second output measurement $y_{2,k}$ are attacked

$$\begin{aligned} \tilde{u}_k &= u_k + a_{1,k} \\ \tilde{y}_k &= \begin{bmatrix} y_{1,k} \\ y_{2,k} + a_{2,k} \end{bmatrix}. \end{aligned}$$

The difference to the system in [7] is that we attack the second output rather than the first output measurement to obtain a bounded optimization problem for the steady state comparison, namely $\ker(G_{ra}) \subseteq \ker(G_{xa})$ as shown in Section 3.4.
This leads to the attacked system

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + B_a a_k \\ \tilde{y} &= Cx_k + D_a a_k \end{aligned}$$

with

$$a_k = \begin{bmatrix} a_{1,k} \\ a_{2,k} \end{bmatrix},\ B_a = \begin{bmatrix} 0.34 & 0 \\ 0.53 & 0 \end{bmatrix} \text{ and } D_a = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

The extended system $\mu_k = [x_k^T, z_k^T]^T$ is then obtained as

$$\mu_{k+1} = \underbrace{\begin{bmatrix} 0.6597 & 0.5335 & -0.0189 & -0.1125 \\ -0.5335 & 0.1262 & -0.0297 & -0.1764 \\ 0.3648 & 0.2667 & 0.2760 & 0.1543 \\ -0.3124 & 0.0846 & -0.2507 & -0.1347 \end{bmatrix}}_{A_e} \mu_k + \underbrace{\begin{bmatrix} 0.3403 & 0 \\ 0.5335 & 0 \\ 0 & 0.2667 \\ 0 & 0.0846 \end{bmatrix}}_{B_e} a_k$$

$$r_k = \underbrace{\begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}}_{C_e} \mu_k + \underbrace{\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}}_{D_e} a_k.$$

One can directly observe that $\ker(D_e) \nsubseteq \ker(B_e)$, so one has to consider the aftermath of the attack. Therefore $\Delta N = 2$ is used to obtain a bounded optimization problem that leads to a stealthy attack, which lasts for $N = 10$ time steps. Finally the threshold of the stateless detector has to be defined, which is also used to obtain the thresholds for the other detectors. Considering stationary zero mean Gaussian measurement and process noise with a covariance matrix $\Sigma_v = \Sigma_w = 10^{-6} I_2$, we get the threshold

$$J_{th} = 2.2865 \cdot 10^{-4}$$

with (3.1) where we set $k = 4$.

4.1.2 Steady State Influence of the Attack

First the steady state influence under the three different detectors is investigated and the results are presented in Figure 4.1, where β is varied between 0 and 1, while δ is varied between 0 and J_{th}. The simulation confirms that the steady state impact of the attack under a MEWMA detector with constant threshold is the same as the one with a stateless detector with the same threshold, but if a threshold dependent on β is chosen wisely, the impact can decrease (linearly for $\beta^2 J_{th}$) with the forgetting factor (see Section 3.4). For the CUSUM detector the impact decreases with the forgetting factor, but this is independent of the threshold chosen. Therefore the simulation shows that choosing either a MEWMA or CUSUM detector can decrease the steady state impact of an attack when the forgetting factor and threshold are chosen appropriately.

4.1.3 Impact on the Whole Trajectory and the Final State

After looking at the steady state both CUSUM and MEWMA seem to have a more restricting influence than the stateless detector on the attacks impact. Therefore the impact of a time-limited attack on the whole trajectory and the final state are investigated, respectively. Here the convex optimization problems (3.8) and (3.9) as well as the CUSUM reformulation (3.13) are used to obtain a numerical solution.
Beginning with the impact on the whole trajectory Figure 4.2 shows the impact under a MEWMA detector with two different thresholds and $\beta \in (0, 1]$. We can see that the MEWMA detector actually benefits the attacker compared to the stateless detector, since we only get a worse impact on the whole trajectory than in the stateless detector case by decreasing the forgetting factor for both

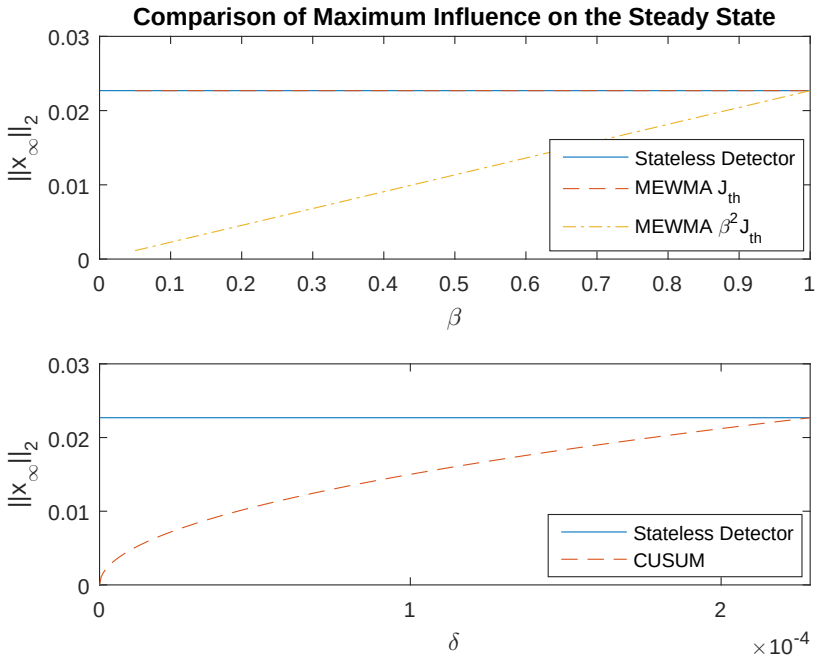

Figure 4.1: Maximum Impact on the State Steady for Different Forgetting Factors

thresholds investigated. While the steady state impact for the threshold $\beta^2 J_{th}$ is decreasing linearly in β it is increasing linearly in β here and it reaches 190% of the impact of the stateless detector with the threshold $\beta^2 J_{th}$ and $\beta = 0.05$. The impact on the system with the stateless detectors threshold is even worse, where it reaches ca. 3800% of the impact of the system under a stateless detector with $\beta = 0.05$.
Figure 4.3 shows the impact on the final state under a MEWMA detector and also here one can see that the detector benefits the attacker, except for the threshold $\beta^2 J_{th}$ and $\beta \geq 0.55$. For these values it seems like the MEWMA detector improves the attack detection. In Figure 4.3 only the impact on the final state $||x_N^*||_\infty$ is plotted and not the impact on the whole trajectory during the attack. Figure 4.4 shows what happens if we solve $\max ||x_N||_\infty$ but look at the maximum impact on $||x||_\infty$. Here we can see that the actual attack impact on the whole trajectory for the attacks that maximize the impact on the final state is always worse than the impact with a stateless detector even for $\beta \geq 0.55$. That the MEWMA detector performs worse than the stateless detector can be explained by its lowpass behavior (see Section 5.1.1).
Let us look at the CUSUM detector now. Figure 4.5 presents the worst-case attack impact on the whole trajectory for different forgetting factors and two different thresholds. One can see that for a constant threshold J_{th}, the attack impact decreases for a forgetting factor $\delta \leq 1.714 \cdot 10^{-4}$ in comparison to the stateless detector, but increases for $\delta > 1.714 \cdot 10^{-4}$. This is not the case for the threshold $J_{th} - \delta$, here we always obtain a more restricted attack impact than

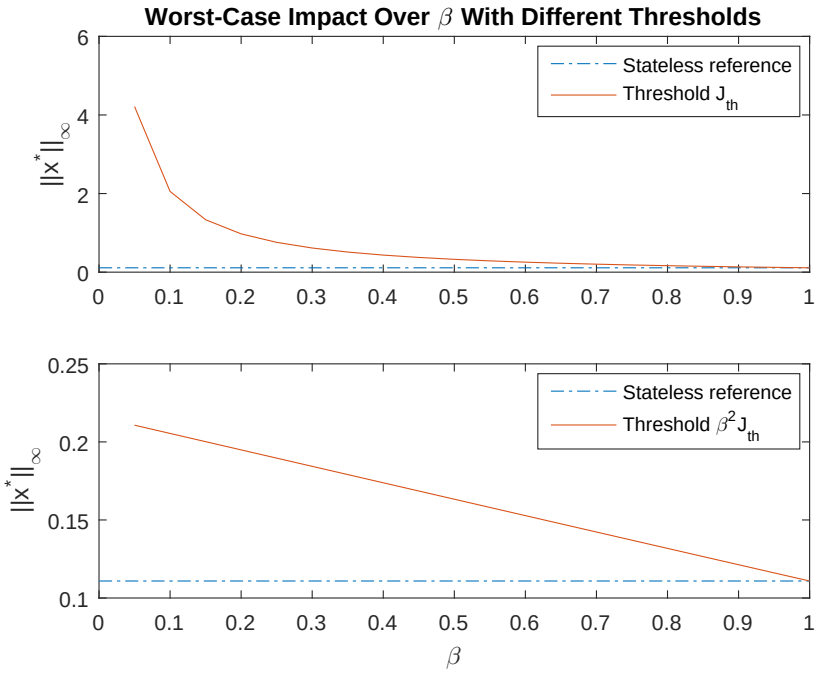

Figure 4.2: Maximum Impact on the Whole Trajectory Using a MEWMA Detector

in the stateless detector case. In equation (3.2) a forgetting factor is proposed to avoid too many false alarm and in this case it results in $\delta \approx 0.914 \cdot 10^{-4}$. In Figure 4.5 we can see that this results in ca. 79.8% of the impact of the stateless detector case for J_{th} and in ca. 74% of the impact in of the stateless detector case for $J_{th} - \delta$.
The same results are obtained for an attack that maximizes the final state impact (see Figure 4.6), where we get an increase in the attack impact for $\delta > 1.206 \cdot 10^{-4}$ when we use the threshold of the stateless detector. For the forgetting factor proposed in (3.2) we get ca. 92.7% of the impact in the stateless detector case for J_{th} and ca. 83% of the impact in the stateless detector case for $J_{th} - \delta$. Therefore δ proposed in (3.2) is a good choice, since it reduces the attack impact compared to the stateless detector case in both worst-case attack scenarios and should avoid too many false alarms. If we look at the impact on the whole trajectory of an attack designed to maximize the impact on $||x_N||_\infty$, the worst-case impact for the threshold J_{th} looks similar to the upper graph in Figure 4.6, while in the case of a threshold $J_{th} - \delta$ we get a slightly higher impact (maximal 5% higher) compared to the stateless detector for $1.184 \cdot 10^{-4} \leq \delta < J_{th}$.

4.1.4 Impact of Stealthy Bang-Bang Attacks

In the this section for the simple simulation example the stealthy bang-bang attack is investigated and its impact on the final state.
Firstly notice that D_e has not full row rank, hence the D_c matrix is not sur-

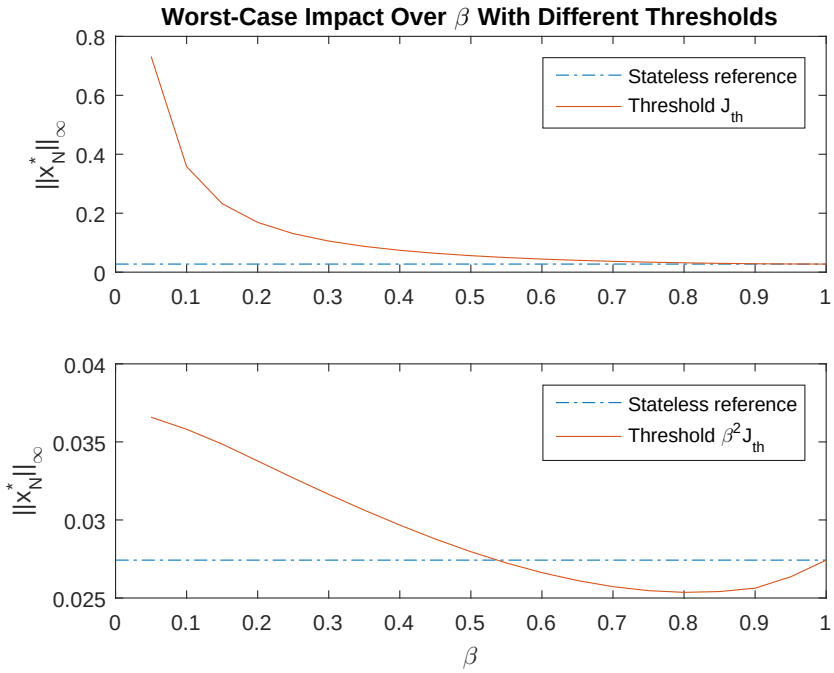

Figure 4.3: Maximum Impact on the Final State Using a MEWMA Detector

jective and therefore the bang-bang attack based on the residuals described in Section 3.7 can not be investigated here.
To investigate the impact of the stealthy bang-bang attack for different detectors the upper bound on the attack c_a has to be determined for different thresholds and forgetting factors. Recall that a way to find an upper bound for the stateless and MEWMA detector is proposed in Section 3.6, but it is more difficult for the CUSUM detector due to its nonlinearity. Therefore we only use the proposed threshold on the residual for a CUSUM detector (see Section 3.6)

$$||r_k||_2 \leq \sqrt{\frac{J_D}{N} + \delta}$$

in this investigation.
The worst-case impact on the final state is shown in Figure 4.7. In the upper graph we can see the worst-case impact for the MEWMA detector with different forgetting factors and two different thresholds and it shows that the MEWMA detector is only reducing the impact of the stealthy bang-bang attack compared to the stateless detector with a threshold of $\beta^2 J_{th}$. The lower graph shows the impact with a CUSUM detector and here the CUSUM detector also reduces the attack in almost every case compared to the stateless detector, except for $\frac{J_{th}}{N} + \delta \geq J_{th}$ in case of the threshold J_{th}.
Lastly we want to show that the upper bound c_a is conservative and Figure 4.8 illustrates this fact for a bang-bang attack using a CUSUM detector with threshold J_{th}. The upper limit for the attack is estimated as $c_a = 0.0052$, which should lead to $||r_k||_2^2 \leq 1.1423 \cdot 10^{-4}$. Figure 4.8 shows that $||r_k||_2^2 \leq 1.7902 \cdot 10^{-5}$, hence

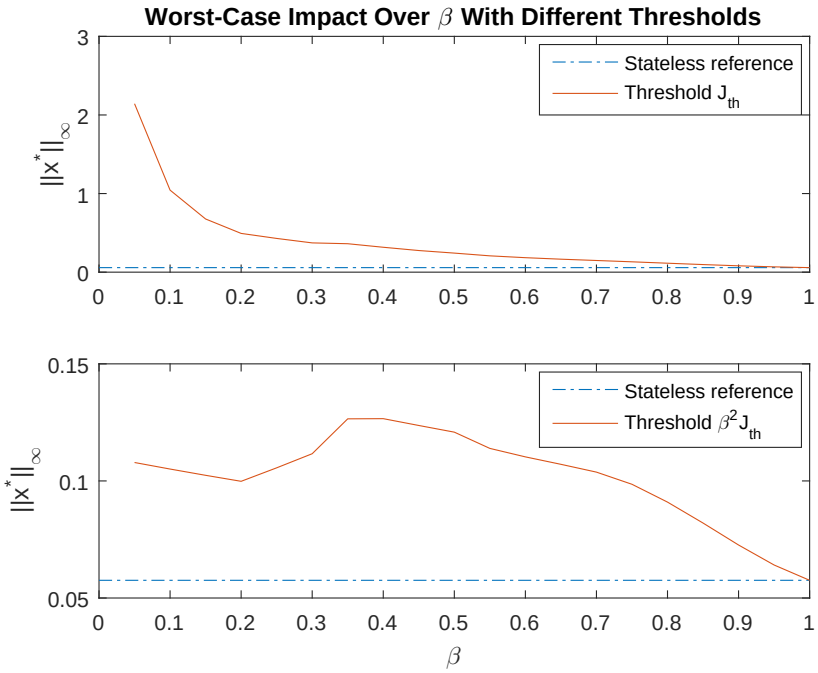

Figure 4.4: Maximum Impact on the Whole Trajectory Of Attacks on the Final State Using a MEWMA Detector

the simulation confirms that the upper bound is conservative, because the residual reaches only 15.67% of the upper bound we aimed at. So the upper bound for the attack can definitely be chosen higher than 0.0052. But it also shows that the attack is stealthy, because $||r_k||_2^2 \leq \delta \approx 9.14 \cdot 10^{-5}$, which results in $S_k = 0 \, \forall k$ and the maximum attack impact is $||x_N^*||_\infty = 0.0066$.

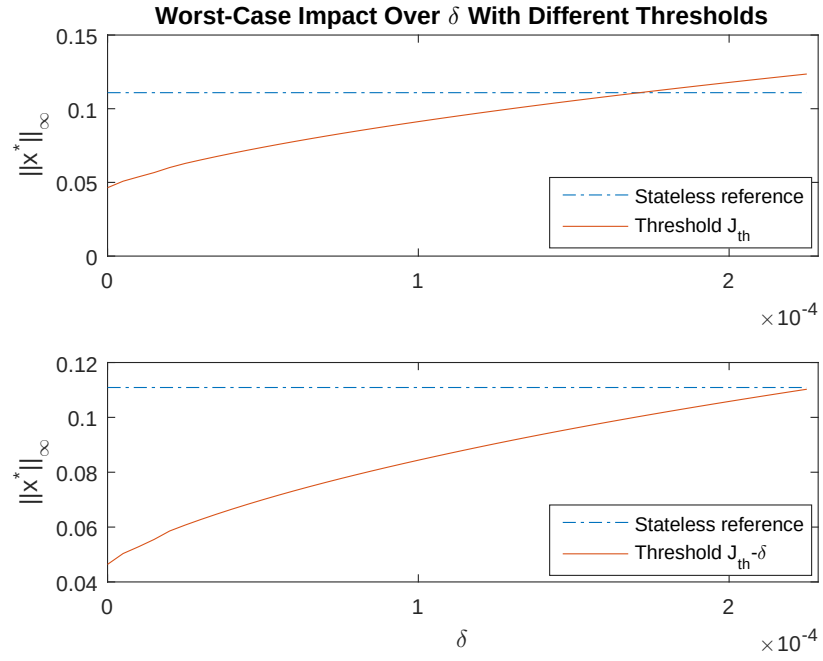

Figure 4.5: Maximum Impact on the Whole Trajectory Using a CUSUM Detector

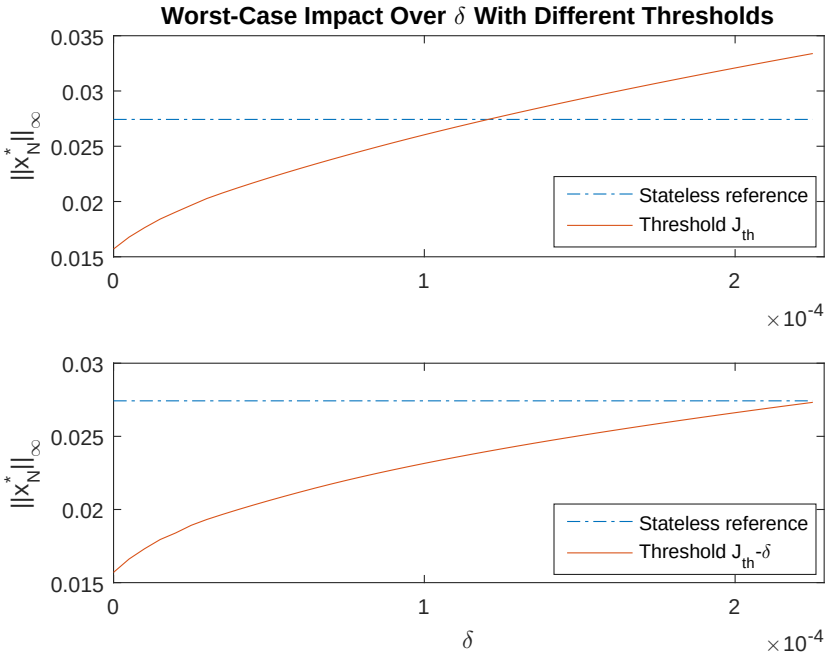

Figure 4.6: Maximum Impact on the Final State Using a CUSUM Detector

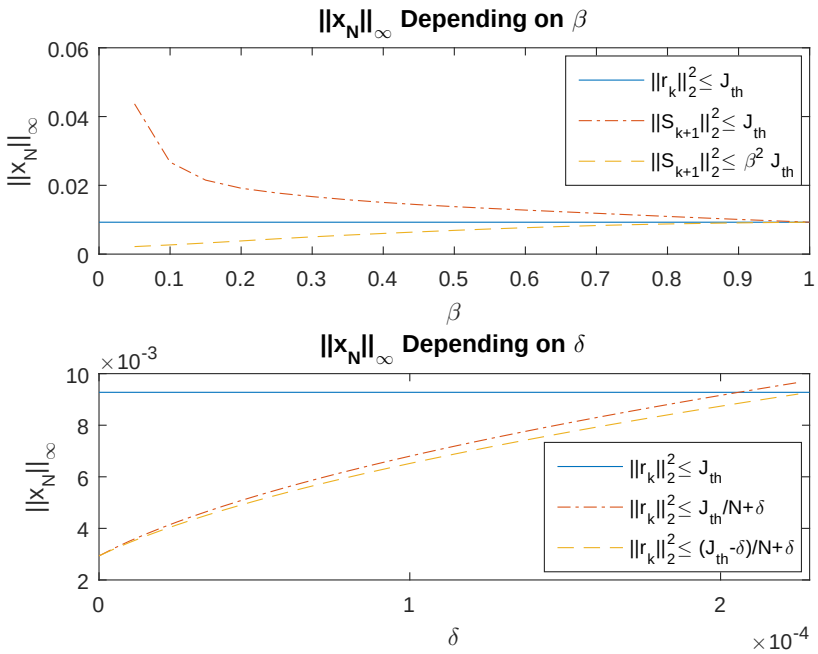

Figure 4.7: Maximum Impact on the Final State Using a Bang-Bang Attack

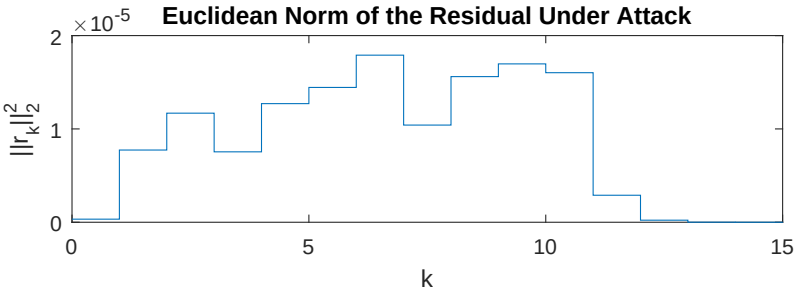

Figure 4.8: Maximum Impact on the Final State Using a Bang-Bang Attack

4.2 Quadruple Tank Process

Now worst-case attack impacts on a realistic process are analyzed under the three detectors. The quadruple tank process is a nonlinear plant, while the attacks aim to disturb the steady state of the system and are defined using a linearized model. Here we characterize attacks and see how each detector restricts their impact in a realistic situation and if they are still undetected. Furthermore data from an experiment with a real quadruple tank process is presented in the end of this section.

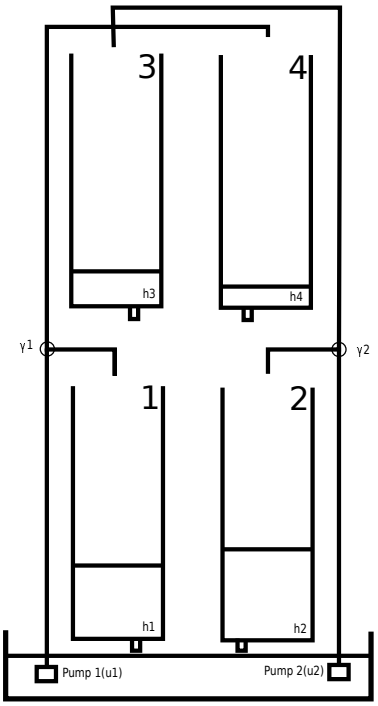

Figure 4.9: Structure of the Quadruple Tank Process

4.2.1 Model of the Quadruple Tank Process

The structure of the quadruple tank process is depicted in Figure 4.9 and it is described by the nonlinear state equations

$$\begin{aligned}
\dot{h}_1 &= -\frac{a_1}{A_1}\sqrt{2gh_1} + \frac{a_3}{A_1}\sqrt{2gh_3} + \frac{\gamma_1 k_1}{A_1}u_1 \\
\dot{h}_2 &= -\frac{a_2}{A_2}\sqrt{2gh_2} + \frac{a_4}{A_2}\sqrt{2gh_4} + \frac{\gamma_2 k_2}{A_2}u_2 \\
\dot{h}_3 &= -\frac{a_3}{A_3}\sqrt{2gh_3} + \frac{(1-\gamma_2)k_2}{A_3}u_2 \\
\dot{h}_4 &= -\frac{a_4}{A_4}\sqrt{2gh_4} + \frac{(1-\gamma_1)k_1}{A_4}u_1 \\
y &= \begin{bmatrix} k_c & 0 \\ 0 & k_c \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix}
\end{aligned}$$

where h_i is the height of each tank, A_i is the cross-section area of each tank, a_i the cross-section area of the outlet holes, g the gravitational acceleration, u_j the voltage applied to the water pumps and k_j are the pump constants. γ_j are the flow ratios of the valves, y represents the output measurements and k_c the measurement constant ($i \in \{1,2,3,4\}$ and $j \in \{1,2\}$).
Linearizing it around a steady state $(h_{1,\infty}, h_{2,\infty}, h_{3,\infty}, h_{4,\infty}, u_{1,\infty}, u_{2,\infty})$ leads to

$$\begin{aligned}
\Delta\dot{x} &= \begin{bmatrix}
-\frac{a_1}{A_1}\sqrt{\frac{g}{2h_{1,\infty}}} & 0 & \frac{a_3}{A_1}\sqrt{\frac{g}{2h_{3,\infty}}} & 0 \\
0 & -\frac{a_2}{A_2}\sqrt{\frac{g}{2h_{2,\infty}}} & 0 & \frac{a_4}{A_2}\sqrt{\frac{g}{2h_{4,\infty}}} \\
0 & 0 & -\frac{a_3}{A_3}\sqrt{\frac{g}{2h_{3,\infty}}} & 0 \\
0 & 0 & 0 & -\frac{a_4}{A_4}\sqrt{\frac{g}{2h_{4,\infty}}}
\end{bmatrix} \Delta x \\
&\quad + \begin{bmatrix}
\frac{\gamma_1 k_1}{A_1} & 0 \\
0 & \frac{\gamma_2 k_2}{A_2} \\
0 & \frac{(1-\gamma_2)k_2}{A_3} \\
\frac{(1-\gamma_1)k_1}{A_4} & 0
\end{bmatrix} \Delta u \\
\Delta y &= \begin{bmatrix} k_c & 0 & 0 & 0 \\ 0 & k_c & 0 & 0 \end{bmatrix} \Delta x
\end{aligned}$$

with $\Delta x_i(t) = x_i(t) - h_{i,\infty}$, $x(t) = [h_1(t), h_2(t), h_3(t), h_4(t)]^T$, $\Delta u_j(t) = u_j(t) - u_{j,\infty}$ and $\Delta y_j(t) = y_j(t) - y_{j,\infty}$. The steady state can be obtained by

$$\begin{aligned}
h_{1,\infty} &= \frac{1}{2g}\left(\frac{(1-\gamma_2)k_2 u_{2,\infty} + \gamma_1 k_1 u_{1,\infty}}{a_1}\right)^2 \\
h_{2,\infty} &= \frac{1}{2g}\left(\frac{(1-\gamma_1)k_1 u_{1,\infty} + \gamma_2 k_2 u_{2,\infty}}{a_2}\right)^2 \\
h_{3,\infty} &= \frac{1}{2g}\left(\frac{(1-\gamma_2)k_2 u_{2,\infty}}{a_3}\right)^2 \\
h_{4,\infty} &= \frac{1}{2g}\left(\frac{(1-\gamma_1)k_1 u_{1,\infty}}{a_4}\right)^2
\end{aligned}$$

where $u_{j,\infty} \in [0, 15]\,\mathrm{V}$ due to the saturation of the pump voltage and $h_{i,\infty} \in [0, 30]\,\mathrm{cm}$, since 30 cm is the maximum height of each tank. To control the system in steady state a linear-quadratic-Gaussian controller is used.
The attacker has the ability to change the first pump voltage Δu_1 and the first measurement sensor Δy_1, so in the discretized system this means

$$a_k = \begin{bmatrix} a_{1,k} \\ a_{2,k} \end{bmatrix}, \ \Delta\tilde{u}_k = \begin{bmatrix} \Delta u_{1,k} + a_{1,k} \\ \Delta u_{2,k} \end{bmatrix} \text{ and } \Delta\tilde{y}_k = \begin{bmatrix} \Delta y_{1,k} + a_{2,k} \\ \Delta y_{2,k} \end{bmatrix}.$$

4.2.2 Simulations

The water tank is modelled as a continuous nonlinear plant in the simulation, but a discrete linear controller is used to control the tanks around their operating point. Therefore measurement signals are sampled and the controller sends step signals to the continuous plant. This simulation is used to examine the effect of the worst-case attack impact under the detectors with a realistic model.

Table 4.1: Simulation Parameters

$A_i[\mathrm{cm}^2]$	$a_1[\mathrm{cm}^2]$	$a_2[\mathrm{cm}^2]$	$a_3[\mathrm{cm}^2]$	$a_4[\mathrm{cm}^2]$
15.179	0.1678	0.1542	0.1591	0.1685
$k_1[\frac{\mathrm{cm}^3}{\mathrm{Vs}}]$	$k_2[\frac{\mathrm{cm}^3}{\mathrm{Vs}}]$	γ_i	$k_c[\frac{\mathrm{V}}{\mathrm{cm}}]$	$g[\frac{\mathrm{cm}^2}{\mathrm{s}}]$
4.32	3.74	0.625	0.2	981

For the simulation of the quadruple tank process the parameters in Table 4.1 are used. The steady state pump voltages are set to $u_{j,\infty} = 6\,\mathrm{V}$ and this results in the operating points

$$\begin{aligned} h_{1,\infty} &= 10.9677\,\mathrm{cm} \\ h_{2,\infty} &= 12.0858\,\mathrm{cm} \\ h_{3,\infty} &= 1.4258\,\mathrm{cm} \\ h_{4,\infty} &= 1.6960\,\mathrm{cm} \end{aligned}$$

The measurement and process noise are assumed to be stationary zero mean Gaussian random processes with covariance matrices $\Sigma_v = 0.0025 I_2$ and $\Sigma_w = 0.01 I_4$, respectively, and the system is sampled with $T_s = 2\,\mathrm{s}$.
First we investigate the numerically obtained worst-case attack impact on the quadruple tank system under detectors with fixed forgetting factors and thresholds to get an impression how the attacks behave. The threshold for the stateless detector and the forgetting factor for the CUSUM are then given by

$$\begin{aligned} J_{th} &= 0.0265\,\mathrm{V}^2 \\ \delta &= 0.0106\,\mathrm{V}^2 \end{aligned}$$

according to equations (3.1) with $k = 4$ and (3.2), respectively. For the MEWMA detector $\beta = 0.5$ is chosen as the forgetting factor.
The attacks last $N = 5$ time steps and also consider their aftermath for $\Delta N = 5$ time steps. Table 4.2 shows the estimated worst-case impact on the whole trajectory (Problem (3.8)) and the final state (Problem (3.9)) under the detectors with the thresholds as defined in the brackets. One can see that the attack

Table 4.2: Attack Impact of Numerical Attacks

	Stateless Detector	MEWMA($\beta^2 J_{th}$)	CUSUM($J_{th} - \delta$)
$\|\|x^*\|\|_\infty$	25.8773 cm	30.5214 cm	20.5981 cm
$\|\|x_N^*\|\|_\infty$	0.9425 cm	0.8643 cm	0.8831 cm

impact on the whole trajectory is huge for all detectors, since it is close to the maximum height of the tanks. Recall that the attack lasts only for $NT_s = 10\,\text{s}$ and is still able to change the water level by at least 20 cm. This behavior seems odd for a realistic process with a limitation on the pump voltage. Therefore we investigate the stealthy attack on the whole trajectory under a CUSUM detector, which is also listed in Table 4.2 and firstly look at the attack signal $a_{1,k}$ on the control input u_1, where the attack starts at 20 s (see Figure 4.10). We

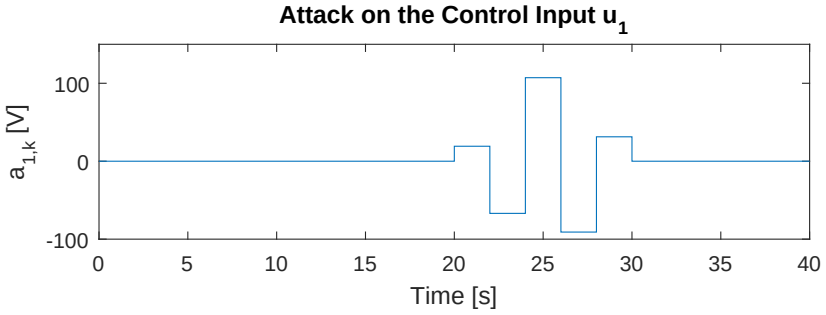

Figure 4.10: $a_{1,k}$ over time

get that $|a_{1,k}| \leq 107.1506\,\text{V}\,\forall k$ so that the controller saturation of 0 V or 15 V is definitely reached and the attack gets cut off. This results in a deviation from the original operating point so that the linear system model used by the detector does not fit the actual nonlinear model of the system any more and the attack gets detected by every detector due to the differences of the nonlinear continuous-time model of the quadruple tank system and the linearized discrete-time model used by the detector (see Figure 4.11). Although this example only illustrates the case for an attack on the whole trajectory under a CUSUM detector, the worst-case attacks both on the whole trajectory and the final state under different detectors show the same behavior.
This illustrates a limitation of the numerically obtained worst-case attacks of Section 3.5. These worst-case attacks are only characterized for linear systems, so that they cannot be used to examine the resilience of the detectors against attacks with a realistic nonlinear system, since the attacks always trigger the alarm and are therefore not stealthy.
Now we want to investigate the worst-case impact of the stealthy bang-bang attacks, that last for $N = 20$ time steps, on the quadruple tank process in the same operating point. In the simple simulation example it has been shown that the thresholds $J_D = \beta^2 J_{th}$ and $J_D = J_{th} - \delta$ are the preferable choice for the MEWMA and CUSUM detectors, respectively, so only these two thresholds will be looked at for the bang-bang attacks. Recall again that the estimation of an upper bound for the attacks under a CUSUM detector is not straight forward,

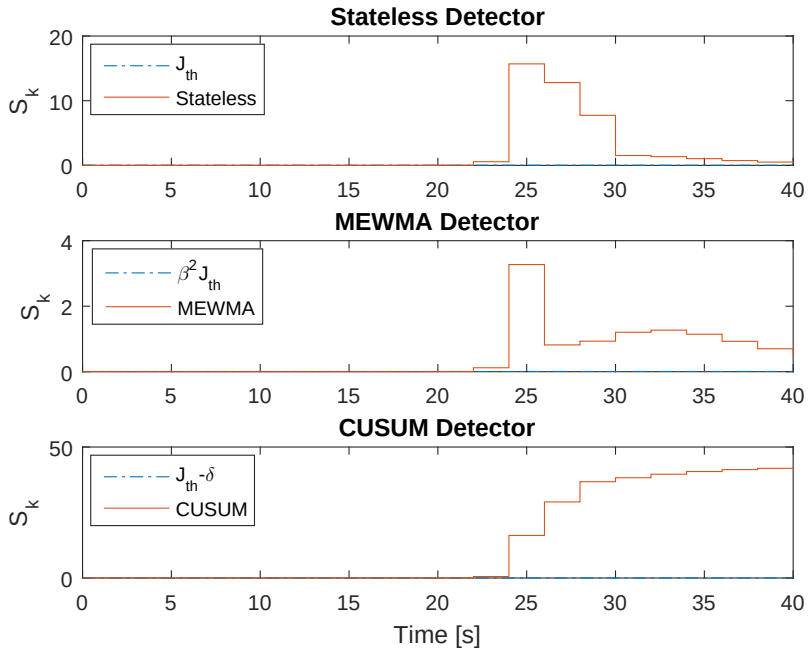

Figure 4.11: Response of the Anomaly Detectors on the Attack

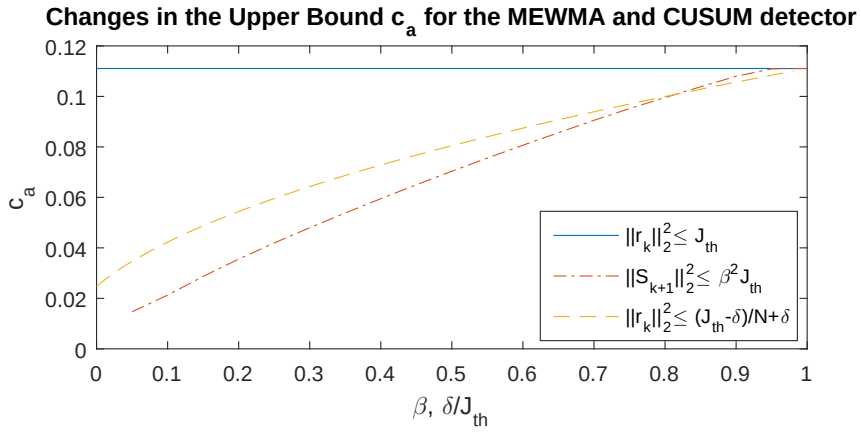

Figure 4.12: Change of c_a over β and δ/J_{th}

so the upper bound proposed in Section 3.6

$$||r_k||_2^2 \leq \frac{J_D}{N} + \delta = \frac{J_{th} - \delta}{N} + \delta$$

is used to characterize a stealthy bang-bang attack under a CUSUM detector. Figure 4.12 displays the change in the upper bound on the attack for the three detectors and one can see that the MEWMA detector with threshold $\beta^2 J_{th}$ actually restricts the attack more for $\beta \leq 0.81$ than the CUSUM detector for

$\delta \leq 0.81 J_{th}$. Furthermore the upper bounds are all smaller than 0.1111 V, hence the attacks will not exceed the controller saturation.
Finally we will examine a stealthy bang-bang attack on the nonlinear plant under a CUSUM detector and see how the detectors behave. With $\delta = 0.0106\,\mathrm{V}^2$ and $J_{th} = 0.0265\,\mathrm{V}^2$ we get the upper bound $c_a = 0.0728\,\mathrm{V}$ on the attack. Since the simple simulation example showed that the calculation of the upper bound is conservative, we can use up to 2.46 times the calculated value for c_a and still remain undetected in this case, hence $c_a = 0.1791\,\mathrm{V}$ leads to a stealthy attack with an estimated impact of $||x_N^*||_\infty = 0.3640\,\mathrm{cm}$ on Tank 1. The actual impact is approximately 0.353 cm and that shows that the estimated impact on the linear system is close to the actual impact on the nonlinear system. Figure 4.13 illustrates the behavior of the anomaly detectors and the attack is not detected by neither the stateless nor the CUSUM detector, but the MEWMA detector with $\beta = 0.5$ is able to detect the attack, which corresponds to the findings of Figure 4.12.

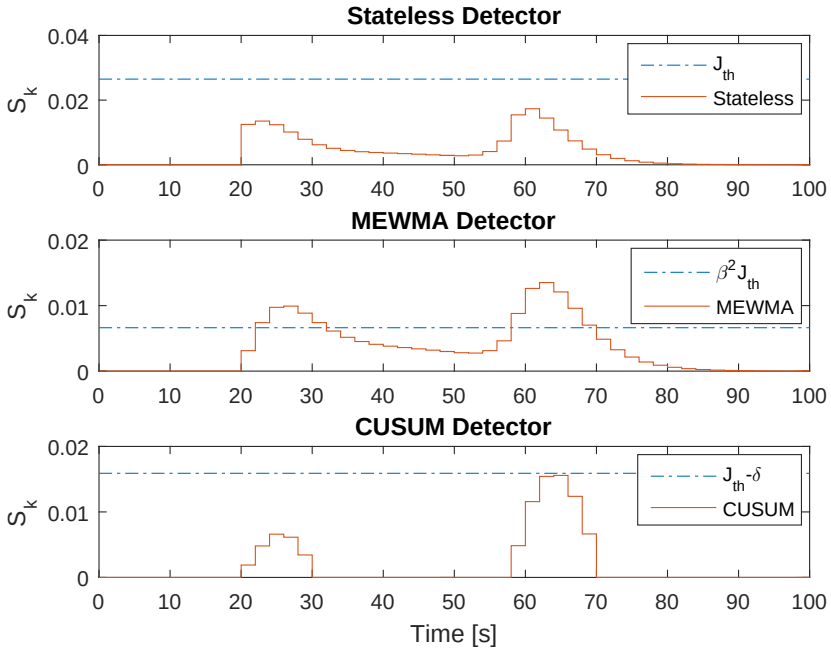

Figure 4.13: Detectors under a Stealthy Bang-Bang Attack

4.2.3 Experimental Results

First it is pointed out that the process is a continuous system and also the controller and Kalman filter are designed for the continuous system. Furthermore there is no network transmitting the data and therefore the system does not fit in the networked control system framework we used previously. Since the process is controlled with a computer the data has to be sampled and the sampling time is actually $T_s = 0.05\,\mathrm{s}$, so it still fits in the discrete control system

framework.
For the real quadruple tank process the parameters are identified as given in Table 4.3. The steady state voltage is again set to $u_{j,\infty} = 6V$, which leads to

Table 4.3: Identified Parameters

$A_i[\mathrm{cm}^2]$	$a_1[\mathrm{cm}^2]$	$a_2[\mathrm{cm}^2]$	$a_3[\mathrm{cm}^2]$	$a_4[\mathrm{cm}^2]$
15.52	0.1448	0.1499	0.1785	0.1737
$k_1[\frac{\mathrm{cm}^3}{\mathrm{Vs}}]$	$k_2[\frac{\mathrm{cm}^3}{\mathrm{Vs}}]$	γ_1	γ_2	$k_c[\frac{\mathrm{V}}{\mathrm{cm}}]$
4.0470	4.3167	0.3927	0.41397	0.2

the steady states

$$h_{1,\infty} = 14.8473\,\mathrm{cm}$$
$$h_{2,\infty} = 14.7130\,\mathrm{cm}$$
$$h_{3,\infty} = 3.6862\,\mathrm{cm}$$
$$h_{4,\infty} = 3.6730\,\mathrm{cm}.$$

One has to keep in mind that these values are derived from an analytical system model that is not representing the real system with complete accuracy. Therefore the true steady states deviate from analytical ones and in the experiments the deviations are around 0 cm to 0.5 cm.
By assuming the measurement and process noise are stationary zero mean Gaussian random processes with a covariance matrix of $\Sigma_v = 0.0025 I_2$ and $\Sigma_w = 0.01 I_4$, respectively, a linear-quadratic Gaussian controller is designed to control the system around its operating point.
The simulation of the quadruple tank has shown that the attacks derived numerically are strong and often hit the controller saturation, so to prevent the real system from damage only a stealthy bang-bang attack is investigated here. The attacker assumes a sampling time of $T_s = 2\,\mathrm{s}$ and chooses $N = 50$, which results in an attack duration of 100 s. The attacker considers a stateless detector with $J_{th} = 0.0292\,\mathrm{V}^2$ and the upper bound of the attack is estimated as $c_a = 0.1072\,\mathrm{V}$. But since the simulation sections showed that it is a conservative bound, we use 1.5 times the bound, so $c_a = 0.1608\,\mathrm{V}$ and the absolute maximum impact is derived as 0.3746 cm, where Tank 2 is the target of the attack.
The heights of the tanks 1 and 2 before and during the attack are shown in Figure 4.14. The attack starts at $30\,s$, where $h_2(30\,\mathrm{s}) \approx 14.9\,\mathrm{cm}$ and increases the water level to $h_2(130\,\mathrm{s}) \approx 15.22\,\mathrm{cm}$, which results in an absolute value of the attack impact of 0.32 cm. This shows that the calculation of the attack impact for a stealthy bang-bang attack is quite accurate even for the real system, since there is only a difference of approximately 0.05 cm. But as one can see the attack leads to an even higher level in Tank 2 ($h_2(147.3\,\mathrm{s}) \approx 15.39\,\mathrm{cm}$) before the controller starts regulating the height back to the nominal water level. Not only the water level in Tank 2 is affected but also the level in Tank 1. The level at the beginning of the attack is $h_1(30\,\mathrm{s}) \approx 15.45\,\mathrm{cm}$ and diminishes to $h_1(130\,\mathrm{s}) \approx 15.1\,\mathrm{cm}$, which results in an absolute value of the attack impact on Tank 1 of 0.35 cm, that is even higher than the impact on the target Tank 2 during the attack.
The stealthiness of the attack is now investigated under different anomaly detectors, where we choose the forgetting factors $\beta = 0.5$ and $\delta = 0.0117$ as proposed

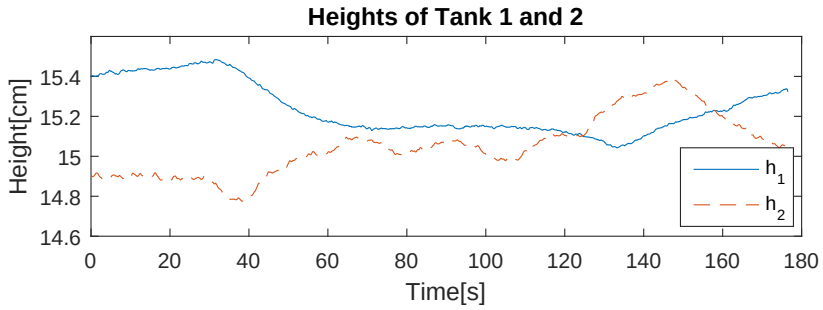

Figure 4.14: Heights of the Watertank under Attack

in (3.2) and thresholds $\beta^2 J_{th}$ and $J_{th} - \delta$ for the MEWMA and CUSUM detector, respectively. The respective thresholds for the MEWMA and CUSUM detector have been chosen, because these limit the attacker the most according to the simulations.

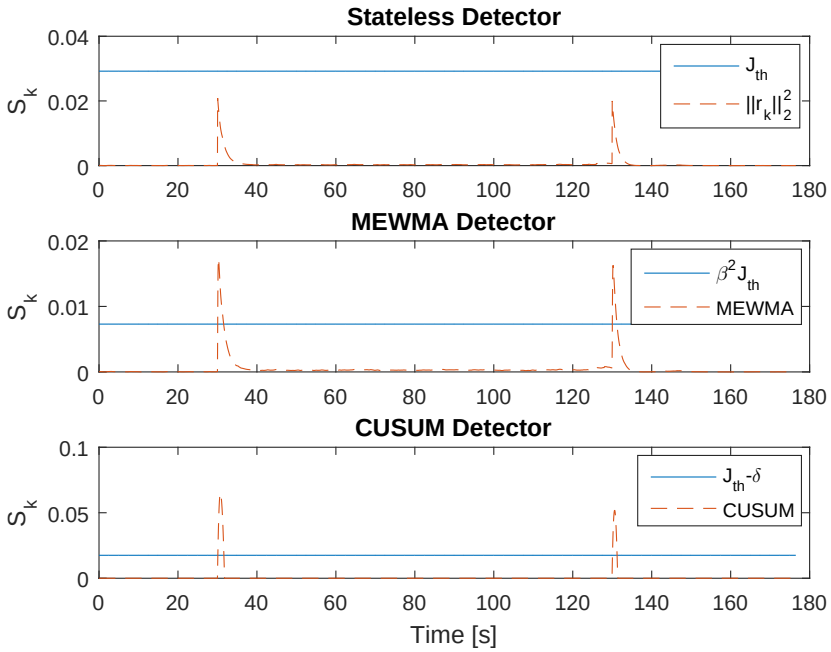

Figure 4.15: Anomaly Detectors of the Real Process

The stateless anomaly detector is not able to detect the attack, while both the MEWMA and CUSUM detector detect the attack (see Figure 4.15). This shows not only that the attack is stealthy for the detector it is designed for but also again that the upper bound is conservative, since the attack uses a higher upper bound than calculated and still remains undetected by the stateless detector. Furthermore it demonstrates also that the MEWMA and CUSUM detector with their respective thresholds are able to detect the attack and therefore put more

restrictions on the adversary. It should also be mentioned that if we use the threshold J_{th} for the MEWMA and CUSUM detector, MEWMA is not able to detect the attack, while CUSUM still detects it. This also corresponds to the simulation results, where this threshold increased the worst-case attack impact for the MEWMA detector but still decreased it for the CUSUM detector (see Figure 4.7).

4.3 Summary

Both simulations and experiments are presented in this chapter to investigate the influence of the detectors on the attack impact. A simple simulation example is used to investigate the general behavior of the detectors and we show that the CUSUM detector is able to diminish the attack impact compared to the stateless detector. The MEWMA detector benefits the attacker in certain cases compared to the stateless detector. In the more sophisticated quadruple tank simulation it is demonstrated that the undetectable worst-case attacks designed for a linearized system can be detected if applied to the nonlinear system due to the nonlinearities such as the controller input saturation. Furthermore experiments on an actual quadruple tank process are performed, where a stealthy bang-bang attack is used to evaluate the detectors.
The results of this chapter are discussed in the following chapter.

Chapter 5

Discussion and Conclusion

In this chapter the results are discussed and the thesis is concluded. Therefore we start with the comparison of the anomaly detectors, continue with the influence of the forgetting factor on the number of false alarms. We present and discuss some peculiarities in the attack and residual signals. Furthermore the results obtained in our work are related to the results in [1]. Lastly we conclude our work and suggest some topics for future work.

5.1 Discussion

5.1.1 Comparison of the Anomaly Detectors

Here the anomaly detectors are compared and a recommendation is given, which one should be used to restrict an adversary the most in all cases considered and which thresholds are preferable for each detector.
The simulation and theory of the worst-case *steady state* impact show that the CUSUM detector is restricting the adversary more compared to the stateless detector independent of the threshold chosen, since it only depends on the forgetting factor δ. The MEWMA detector on the other hand only limits the attacker more compared to the stateless detector case if the threshold $\beta^2 J_{th}$ is chosen. Then the attack impact decreases linearly in β, while the impact under the CUSUM detector decreases according to $\sqrt{\delta}$.
Looking at the worst-case attack impact on the whole trajectory or the final state under the three detectors characterized by the optimization problems (3.8) and (3.9) we see that the MEWMA detector actually benefits the attack impact in the simple simulation example for both thresholds J_{th} and $\beta^2 J_{th}$ in both optimization problems compared to the stateless detector.
CUSUM on the other hand diminishes the attack impact on the whole trajectory compared to the stateless detector for almost all δ and both thresholds. That CUSUM is not able to diminish the attack impact for all δ with J_{th} comes from the fact that a bigger δ means that the system forgets more of the current residual r_k. This benefits the adversary because the detector starts to forget the adversaries actions for a big enough δ if the threshold is constant. This is not the case for the threshold $J_{th} - \delta$, since the more we forget here the less deviation from the initial value of 0 is allowed. Similar results for the impact on the final state are obtained for the CUSUM detector.

The question why the MEWMA detector actually benefits the attacker in the worst-case attacks naturally arises and two possible reason for this are presented in the following. Recall the MEWMA detector is given by

$$S_{k+1} = \beta r_k + (1 - \beta) S_k$$

with $S_0 = 0$. This structure represents a low-pass filter for each element of the residual signal r_k, so that highly oscillating residual signals are suppressed and as $\beta \to 0$ the oscillating residuals are more and more suppressed (see Bode diagram in Figure 5.1). Therefore an intelligent adversary can exploit this low-pass

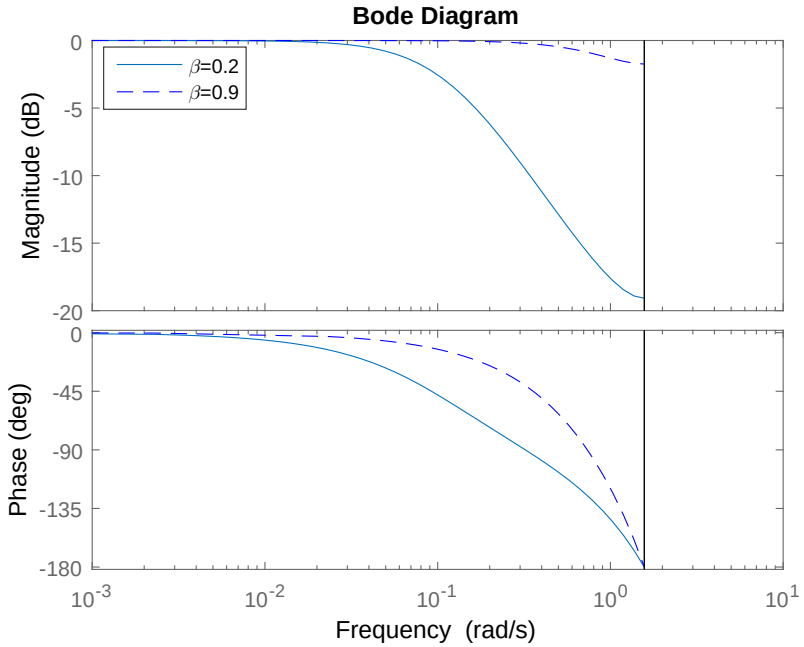

Figure 5.1: Bode Diagram of two EWMA filters with a sampling time of $T_s = 1\,\mathrm{s}$

behavior and create an oscillating residual signal, which leads to a big impact without being detected. Figure 5.2 shows that an attack on the whole trajectory of the simple simulation example under a MEWMA detector with $\beta = 0.2$ and threshold $\beta^2 J_{th}$ leads to a highly oscillating residual, which might exploit the aforementioned low-pass characteristics of the detector. Another reason for the vulnerability of the MEWMA detector to attacks might be the controllability of the extended system. For the simple simulation example the extended system including the plant, the observer and the MEWMA detector is actually controllable with the attack signal a_k as an input, so an intelligent adversary could also be able to take advantage of that and steer the system to an undesirable state while remaining undetected.
However, one has to keep in mind that the worst-case attacks characterized by (3.8) and (3.9) are only for linear systems and lead to detectable attacks in a more sophisticated system as the quadruple tank process simulations show. Hence one should not abandon the MEWMA detector as an attack-resilient

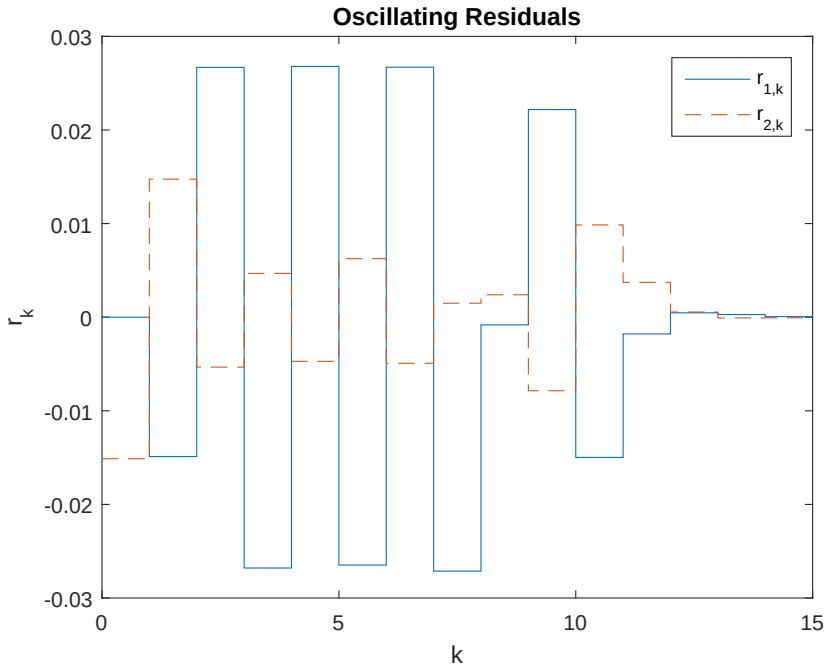

Figure 5.2: Residual Signal of an attack under the MEWMA detector

anomaly detector. Especially in the case of a limited attacker ($||a_k||_2 \leq c_a$) the MEWMA limits the adversary more than the CUSUM detector as shown in the quadruple tank process (see Figure 4.12), if the forgetting factor and threshold are chosen appropriately. However the CUSUM detector has the advantage that it limits the attacker for almost all configurations in the bang-bang attack case, except when δ is chosen too big for the same reasons as mentioned before. Furthermore keep in mind that the upper bounds on the attack c_a proposed are conservative and especially in the case of the CUSUM detector the bound investigated is not necessarily the worst-case bound. Therefore one cannot say in general that the MEWMA detector with the threshold $\beta^2 J_{th}$ is able to restrict the adversary more than the CUSUM detector.
Overall the CUSUM detector is the recommended detector to limit the adversaries worst-case impact on the plant, because it restricts the attacker more compared to the stateless detector in almost every configuration in the attack scenarios investigated. Furthermore the CUSUM detector does not show a strong dependency on the threshold chosen as the MEWMA detector, which can actually benefit the attacker for certain thresholds. Compared to the MEWMA algorithm, CUSUM is also able to limit the attacker in most scenarios investigated and only benefits the attacker for too big forgetting factors, which should not be chosen in a realistic situation. A disadvantage of the CUSUM detector is its nonlinearity, which makes it harder to analyze theoretically, while the MEWMA detector does not suffer this problem.
Furthermore the proposed forgetting factor δ in equation (3.2) is able to limit the adversary attack in all cases considered and is therefore the forgetting factor

to choose, if it proves to avoid too many false alarms as well.
However, we cannot only recommend an anomaly detector but also a way to design the threshold. As it turns out, a threshold, which changes accordingly with the forgetting factor, is more sensitive to attacks, namely $\beta^2 J_{th}$ and $J_{th} - \delta$. These thresholds have shown a better sensitivity towards attacks in the MEWMA and CUSUM detectors than using a constant threshold, which does not incorporate the forgetting factor. One should keep in mind that these thresholds show a higher sensitivity for the attacks, but this might not be true for the noise and could result in more false alarms. Therefore noise sensitivity, which is not investigated in our work, has to be considered as well when one chooses a threshold. Urbina et al. [1] consider the stochastics for the CUSUM and stateless detector and discover a trade-off between the attack impact and the mean-time between false alarms, when one chooses the threshold.

5.1.2 Influence of the Forgetting Factor

Now that we compared the anomaly detectors let us take a look at the role of the forgetting factor. In all the simulations concerning the CUSUM detector the attacker gets more and more restricted if $\delta \to 0$. Therefore one could wrongly assume that $\delta = 0$ is the best configuration for the CUSUM detector, because it leads to an increase in false alarms due to the noise and model uncertainty. To explain this we look at the CUSUM detector with $\delta = 0$

$$S_{k+1} = \sum_{i=0}^{k} ||r_i||_2^2.$$

Generally we have $||r_i||_2^2 \neq 0$ with noise and model uncertainty, so that all the nominal residuals sum up over time and trigger false alarms. This explains why $\mathbb{E}(||r_k||_2^2 - \delta) < 0$ is recommended in Section 2.3. Therefore one has to consider a trade off between the number of false alarms as well as the sensitivity to attacks, when designing the forgetting factor of the anomaly detector.
Not exactly the same argument holds for the MEWMA. If the threshold J_{th} is used, the highest restriction in the attack occurs for $\beta \to 1$, while this also increases the false alarm rate, since we do not forget much of the current residual. If $\beta \to 0$ we forget more of the current residual and therefore decrease the false alarms rate, but it also benefits the attacker, since it improves stealthiness. Clearly for $\beta = 0$ the attacker can have as much impact as he wants, because the current residual is not considered at all.
For the threshold $\beta^2 J_{th}$ the attack impact decreases for $\beta \to 0$ in certain cases, like the steady state impact and for the bang-bang attacks, but for $\beta = 0$ it should increase again due to aforementioned reasons. Since the threshold also decreases with β one can assume that the false alarm rate increases again compared to the threshold J_{th}. Therefore the threshold chosen for the MEWMA detector has to be considered in the trade-off between false-alarm rate and the sensitivity to the attacks, while the CUSUM shows the same behavior for both thresholds.

5.1.3 Peculiarities in the Attack and Residual Signals

Finally we want to take a look two peculiarities in the attack signals and one peculiarity in the residual signal.
An attack on the final state, when it takes its aftermath ΔN into account, exhibits both peculiarities we want to depict. Therefore an attack on the final state is characterized for the simple simulation example under a MEWMA detector with $\beta = 0.2$ and threshold $\beta^2 J_{th}$, which lasts for $N = 50$ time steps and considers the aftermath for $\Delta N = 3$ time steps to remain undetected. The state and attack trajectory are now investigated(see Figure 5.3). On the one hand we

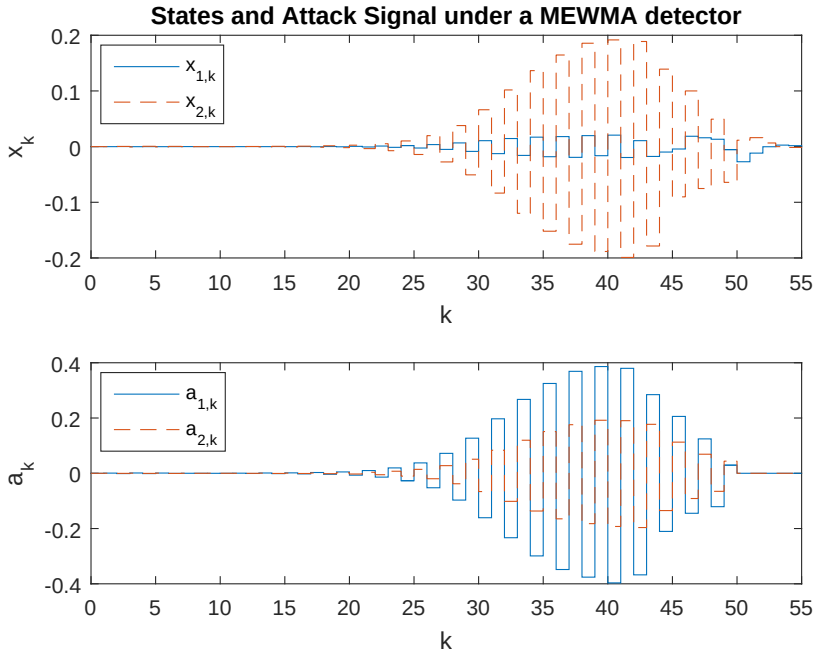

Figure 5.3: Investigation of the attack itself and the state trajectory

want to show that the attack has its maximum impact on the state trajectory before the attacks last hit a_N(see upper graph in Figure 5.3), hence to remain stealthy the adversary has to put some of his time and energy in reversing the attack effects to bring the system back to a believable state and remain undetected. This also corresponds to Figure 4.4 where the impact on the whole trajectory for an attack targeting the final state is depicted. So looking at the attacks impact on the final state does not necessarily say something about the overall impact on the system, when the adversary considers the aftermath of his attack.
On the other hand we want to illustrate that although the attack starts at $k = 0$, $a_k = 0$ for the first time steps (see lower graph in Figure 5.3). Therefore it seems that an optimal attack duration N^* exists when the adversary targets the final state. For $N > N^*$ the adversary will not have a higher impact on the trajectory and N^* depends most probably on the stability of the system(i.e. how fast $A^k \to \mathbf{0}$ for $k \to \infty$), since this determines how fast the effects of the

attack decays. The optimal attack time is highlighted here because it might also be a way to limit the adversary, since the higher N^* the more time it takes the attacker to reach his worst impact on the system, hence there is also more time to detect the attack. This optimal attack time N^* is not further investigated here, but should be looked into more to discover more ways to limit the adversary.
Figure 4.15 illustrates a peculiarity of the residual signal under a stealthy bang-bang attack. We see that the residual exhibits only strong deviations from 0 when the attack starts and ends, but not during the attack. The attack is only detectable in this case directly when it starts or when it ends otherwise the system behavior appears to be nominal. Therefore the detectors should be able to detect quick changes in the residual, which CUSUM and MEWMA are designed for, when you for example recall CUSUM's optimality in change detection.

5.1.4 Comparison to Article [1]

Since [1] also investigates the CUSUM detector, this section will discuss how our work complements and extends the work done in [1].
Let us compare our work and [1] in the following two aspects.

- **Attack Characterization**
  Our work characterizes the worst-case attacks by setting up and solving an optimal control problem, where the attack signals on both the actuator and measurement signals as well as the physical attacks are used as optimization variables.
  The attacks in [1] are characterized in a more intuitive way. Instead of maximizing the state of the system as in our case, the optimization problem in [1] is to maximize the difference between the measurements of the nominal behavior and the systems behavior under attack. The attacks are designed by investigating the conditions of not being detected under the detectors to get an intuitively optimal attack. The attacks consider also the noise in this way, which our attack characterization neglects. Furthermore either actuator or measurement attacks can be designed, but not both together as in our case and the solution does not yield optimal attack vectors a_k but rather the optimally changed measurement or actuator signal by the attacker.
  Both [1] and our work consider a time-limited attacker, which has perfect knowledge of the system and the detector used.
  Therefore our attack model can be seen as an extension to the model used in [1]. The advantage of the attack model in [1] is that it considers the noise processes as well, which are neglected in our model.

- **Evaluation of the Detectors**
  In [1] a new evaluation metric is proposed, where the worst attack impact on the system is plotted over the mean time between false alarms. In their work, the thresholds of the stateless and CUSUM detector are varied to determine the mean-time between false alarms and the attack impact. CUSUM's forgetting factor is fixed during the sweep over the threshold.
  In our work we also consider worst impact of the attack, but we fix the threshold and sweep over the forgetting factor, where the worst attack impact on the stateless detector is used to compare the performance of

the CUSUM detector under different forgetting factors with the one of the stateless detector. Furthermore, the false alarms are only indirectly considered while determining the threshold of the stateless detector.
Despite the difference evaluation metrics the results are in agreement. Both metrics show that the CUSUM is to be preferred over the stateless detector. On the one hand, CUSUM results in smaller worst attack impact than the stateless detector with the same threshold and time between false alarms according to [1]. On the other hand, the attack impact with the CUSUM detector is also reduced if the forgetting factor is chosen appropriately. Therefore our work complements the results of [1], since we look at different variables of the anomaly detectors and obtain similar results for the attack impact under different detectors.

We recommend to combine the results of our work and [1] in future studies to get a more sophisticated analysis of the different detectors presented here. With the novel evaluation metric proposed in [1], we can analyse the worst-case attacks characterized in our work further and also in a more realistic setting where the false alarms are considered.

5.2 Conclusion and Future Work

This section concludes the thesis and proposes future research topics.

5.2.1 Conclusion

In our work three control system anomaly detectors, a stateless detector, a CUSUM detector and a MEWMA detector, are investigated and compared. The ability of each detector to limit the stealthy worst-case attack impact is investigated. To determine the worst-case attack impact, optimal control problems are formulated and solved both analytically and numerically for a simple simulation example as well as a sophisticated quadruple tank process. Furthermore an experiment with a physical quadruple tank process is conducted to observe how the detectors behave in a real-life situation. The CUSUM detector reduces the attack impact for almost every configuration in each scenario investigated compared to the stateless detector and the MEWMA detector. Only if the forgetting factor of the CUSUM detector is chosen too big, it limits the attacker less than the stateless detector. The MEWMA detector has the disadvantage that it can benefit the attacker compared to the other detectors and its sensitivity shows a strong dependency on the threshold chosen in contrary to the CUSUM detector. For that reason the CUSUM detector is the detector to choose, if one wants to limit the attack impact in all cases considered. A way to design the thresholds of anomaly detectors with forgetting factors is also presented, where the thresholds should change appropriately with the forgetting factor.

5.2.2 Future Work

Possible future work is described in the following. One can look more deeply into the three detectors and not only compare their sensitivity to data injection attacks but also to other attacks. Furthermore the sensitivity of the detector to an attack, where we consider the noise should also be investigated more. Here only three anomaly detectors are looked at, so looking into more anomaly detectors is also recommended. Especially investigating linear detectors like the MEWMA detector is recommended, because they can be included in the systems state and have a simpler mathematical treatment than nonlinear detectors. One could also investigate a combination of the CUSUM and MEWMA detector or look into the EWMA detector, where the squared Euclidean norm of the residuals is used as in the CUSUM detector. The stealthy bang-bang attacks offer also a few open research topics, like the estimating a non-conservative upper bound c_a, especially in the CUSUM case.
Furthermore the discovered attack peculiarities, namely the optimal attack time N^* and that the attacker has to reverse its effect on the system to remain undetected, are also worth investigating since they could also lead to more ways to restrict the adversary.

List of Tables

List of Figures

Bibliography

[1] D. Urbina, J. Giraldo, A. A. Cárdenas, J. Valente, M. Faisal, N. O. Tippenhauer, J. Ruths, R. Candell, and H. Sandberg, "Limiting the impact of stealthy attacks on industrial control systems," in *Proceedings of the 23rd ACM Conference on Computer and Communications Security*, October 2016, to be published.

[2] 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. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0005109814004488

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

[4] 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, 2015.

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

[6] 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.

[7] A. Teixeira, K. C. Sou, H. Sandberg, and K. H. Johansson, "Secure control systems a quantitative risk management approach," *IEEE Control Systems*, vol. 35, no. 1, pp. 24–45, 2015. [Online]. Available: http://dx.doi.org/10.1109/MCS.2014.2364709

[8] A. Teixeira, H. Sandberg, and K. H. Johansson, "Strategic stealthy attacks: the output-to-output gain l2-gain," in *Proceedings of the 54th IEEE Conference on Decision and Control (CDC)*, 2015, pp. 2582 – 2587.

[9] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954.

[10] G. Lorden, "Procedures for reacting to a change in distribution," *Ann. Math. Statist.*, vol. 42, no. 6, pp. 1897–1908, 12 1971.

[11] G. V. Moustakides, "Optimal stopping times for detecting changes in distributions," *Ann. Statist.*, vol. 14, no. 4, pp. 1379–1387, 12 1986. [Online]. Available: http://dx.doi.org/10.1214/aos/1176350164

[12] Y. Ritov, "Decision theoretic optimality of the cusum procedure," *Ann. Statist.*, vol. 18, no. 3, pp. 1464–1469, 09 1990. [Online]. Available: http://dx.doi.org/10.1214/aos/1176347761

[13] S. W. Roberts, "Control chart tests based on geometric moving averages," *Technometrics*, vol. 42, no. 1, pp. 97–101, 2000.

[14] D. C. Montgomery, *Introduction to Statistical Quality Control*, 6th ed. Wiley, 2009.

[15] 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.

[16] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*, ser. Prentice-Hall information and system sciences series. Prentice Hall, 2000. [Online]. Available: https://books.google.se/books?id=zNJFAQAAIAAJ

[17] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," nov 2012, version 20121115. [Online]. Available: http://www2.imm.dtu.dk/pubdb/p.php?3274

[18] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[19] J.-B. Hiriart-Urruty, "Global optimality conditions in maximizing a convex quadratic function under convex quadratic constraints," *Journal of Global Optimization*, vol. 21, no. 4, pp. 443–453, 2001. [Online]. Available: http://dx.doi.org/10.1023/A:1012752110010

[20] T. L. Friesz, *Dynamic Optimization and Differential Games*, ser. International Series in Operations Research and Management Science. Springer US, 2010.

[21] 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.

TRITA TRITA-EE 2016:093
ISSN 1653-5146