Data-Driven Methods for the Safety and Security of Constrained Cyber-Physical Systems

Mehran Attar

A Thesis

in

The Department

of

Concordia Institute for Information and Systems Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy (Information and Systems Engineering) at

Concordia University

Montréal, Québec, Canada

February 2025

Concordia University

School of Graduate Studies

This is to certify that the thesis prepared

By: **Mr. Mehran Attar**

Entitled: **Data-Driven Methods for the Safety and Security of Constrained Cyber-Physical Systems**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Information and Systems Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Rabindranath Raut

_____ External Examiner
Dr. Davide M. Raimondo

_____ Arm-Length Examiner
Dr. Shahin Hashtrudi Zad

_____ Examiner
Dr. Rastko Selmic

_____ Examiner
Dr. Arash Mohammadi

_____ Supervisor
Dr. Walter Lucia

Approved by _____
Dr. Farnoosh Naderkhani,
Graduate Program Director

December 4th, 2024

_____
Dr. Mourad Debbabi, Dean,
Gina Cody School of Engineering and Computer Science

Abstract

Data-Driven Methods for the Safety and Security of Constrained Cyber-Physical Systems

Mehran Attar, Ph.D.

Concordia University, 2024

Cyber-Physical Systems (CPSs) are typically defined as physical systems that are integrated with computational and communication capabilities, offering the potential to significantly enhance traditional engineering systems in terms of efficiency, reliability, and performance. However, these enhanced features also introduce potential vulnerabilities to cyber-attacks, as evidenced by the various types of attacks reported against CPSs.

Over the past decade, numerous control solutions have been proposed to detect these attacks and mitigate their impact on CPSs. Most of the proposed solutions are based on the assumption of precise a priori knowledge of the system's dynamical model. However, acquiring an accurate mathematical model can be challenging, particularly when the system's behavior is affected by unknown or uncertain factors, such as disturbances. Consequently, recent efforts have focused on developing data-driven control approaches to safeguard the safety of constrained CPSs against cyber-attacks.

In this thesis, the safety and security problems in constrained Cyber-Physical Systems using data-driven methods are addressed. In the first part of this thesis, we propose an active detection mechanism using the concept of dimensionality reduction, Principal Component Analysis (PCA), and optimal reconstruction for the detection of intelligent and coordinated cyber-attacks. In particular, we design a novel attack detection strategy based on a time-varying (random) encoding mechanism, which encodes the sensor measurements into a random latent space and prevents the attacker from accessing some part of the information.

In the second part of this thesis, we first introduce a data-driven method for computing backward reachable sets and Set-Theoretic Model Predictive Control (ST-MPC) tailored for constrained control systems, subject state, and control input constraints. Then, we use the developed data-driven ST-MPC along with data-driven backward and forward reachability concepts to design data-driven detection and mitigation control architectures aimed at preserving the safety and tracking performance of constrained CPSs against cyber attacks. In particular, in the first architecture, a data-driven anomaly detector based on the forward one-step evolution of the system is designed to detect the presence of attacks on the measurement channel. Moreover, on the plant's side, a data-driven safety verification module is designed to assess whether the received control input ensures a safe evolution of the plant. If necessary, it replaces the networked controller with a local data-driven, set-theoretic model predictive controller, which aims to maintain the plant's trajectory within a predefined safe configuration until normal operation is restored following the attack. In the second architecture, we extend the first solution in order to minimize whenever possible, the tracking performance loss of constrained CPSs in the presence of cyber attacks on the measurement channel. For this purpose, an add-on tracking supervisor module is designed which operates in an open-loop fashion in case of unreliable measurements. Moreover, on the plant side, the safety verification module is enhanced to consider multiple safety equilibrium points which allows to reduce the tracking performance loss in the presence of cyber-attacks on the actuation channel.

Acknowledgments

I would like to express my deepest gratitude to the following individuals, whose support was indispensable not only in completing this research but also in guiding me through my Ph.D. journey.

First and foremost, I wish to extend my deepest gratitude to my supervisor, Dr. Walter Lucia, whose expertise and insight guided me through this research. His enthusiasm for the project, along with his unwavering support, encouragement, and patience, provided me with the direction and confidence to stay on the right path.

Over the past four years, I faced many challenges, and I could not have overcome them without the unwavering support of my family. Words cannot fully express my gratitude for the love, encouragement, and support of my fiancée, Sadaf. I am also deeply thankful for the constant love and guidance of my mother, sister, and brothers throughout this journey. This achievement would not have been possible without them.

I dedicate this work to the cherished memory of my dear brother, Dr. Mojtaba Fayazi, whose constant support and encouragement have always guided me.

Last but not least, I am deeply grateful to my friends and teammates at Concordia University, especially Samaneh, Behrouz, Hamed, Farzaneh, Cristian, Geovana, and Surya, for their invaluable contributions, thoughtful advice, and steadfast support.

Montreal, December 2024 Mehran Attar

Contents

List of Figures

List of Tables

List of Symbols

Symbol	Description
$\mathbb{R}$	The set of real numbers
$\mathbb{R}^n$	The set of real-valued column vectors of dimension n
v_k	Values of v at the discrete sampling time instant $k \in \mathbb{Z}_+$, where $\mathbb{Z}_+ := \{0, 1, \dots\}$
v^T	The transpose of a vector v
M^T	The transpose of a matrix M
O_n	Square matrix $n \times n$ of zeros
I_n	Identity matrix of $n \times n$
$0_{m \times n}$	Rectangular matrix of zeros with dimensions $m \times n$
M^j	j-th power of a square matrix M
$v_k \sim \mathcal{N}(\mu_v, \Sigma_v)$	v is normally distributed with mean μ_v and covariance matrix $\Sigma_v > 0$
M_{*j}	j-th column of $M \in \mathbb{R}^{n \times m}$
$\bar{M} \in \mathbb{R}^{N \times d}$	The normalized matrix $M \in \mathbb{R}^{N \times d}$, where $\bar{M}_{*j} = \frac{M_{*j} - \mu_j}{\Sigma_j}$ for all j, and μ_j, Σ_j are the mean and covariance of M_{*j}
$V^{(N)} \in \mathbb{R}^{m \times N}$	A collection of N samples of v_k
$\|v_k\|$	Euclidean norm of v_k, defined as $\|v_k\| = \sqrt{v_k^T v_k}$
$\|v_k\|_\infty$	Infinite norm of v_k, $\|v_k\|_\infty = \sup_{k \geq 0} \|v_k\|$
$M^\dagger$	Right pseudo-inverse of matrix M

$CP_n(M)$	n-dimensional cross-product of matrix M, defined as $CP_n(M) = [\det(M^{[1]}), \ldots, (-1)^{j+1} \det(M^{[j]}), \ldots, (-1)^{n+1} \det(M^{[n]})]^T$ [3]
$M^{[j]}$	Sub-matrix of $M \in \mathbb{R}^{(n-1)\times(n-1)}$, where the j-th row is removed

List of Abbreviations

CPS:	Cyber Physical System
NCS:	Networked Control System
C-P:	Controller to Plant
P-C:	Plant to Controller
FDI:	False Data Injection
CIA:	Confidentiality, Integrity, and Availability
DoS:	Denial of Service
LTI:	Linear Time Invariant
RCI:	Robust Control Invariant
ROSC:	Robust One Step Controllable
RORS:	Robust One-Step Reachable Set
BRS:	Backward Reachable Set
UIO:	Unknown Input Observer
FP:	False Positive
FN:	False Negative
TP:	True Positive
TN:	True Negative
CSTR:	Continuous-Stirred Tank Reactor
IID:	Independent and Identically Distributed
DoA:	Domain of Attraction
MPC:	Model Predictive Control
ST-MPC:	Set-Theoretic Model Predictive Control

D-STMPC:	Data-Driven Set-Theoretic Model Predictive Control
QP:	Quadratic Programming
EC:	Emergency Controller
TS:	Tracking Supervisor
MAC:	Message Authentication Code
SV:	Safety Verification
PCA:	Principal Component Analysis

Chapter 1

Introduction and Literature Review

1.1 Cyber-Physical Systems

A Cyber-Physical System (CPS) refers to a physical system integrated with communication and control capabilities. CPSs are prevalent in modern society, ranging from small to large-scale applications, including unmanned aerial vehicles, chemical plants, autonomous transportation systems, smart grids, and water distribution networks [4, 5]. CPSs have the potential to significantly enhance traditional engineering systems in terms of efficiency, reliability, and performance. However, these advancements also introduce new vulnerabilities to cyber-attacks that target the cyber infrastructure and communication networks. In the control systems community, CPSs are often modeled as Networked Control Systems (NCS), where adversarial agents can disrupt the closed-loop system's performance by executing cyber-attacks on the communication channels [6, 7]. The control systems community has been actively engaged in addressing the security, safety, and privacy challenges associated with CPSs, as evidenced by numerous studies (see e.g., [8–13] and references therein). Although extensive research has been done to address NCSs problems [14], the study of cyber-attacks affecting NCSs is still a relatively young research area [12]. Cyber-attacks on industrial plants could cause irrecoverable damages with huge financial losses. Several cyber-attacks affecting CPSs have been reported in the last decade, see, e.g., the well-known Stuxnet [15] and the recent Industroyer [16] malware. Due to economic and political interests, the number of cyber threats on physical control systems has increased in recent years,

and more are expected in the near future.

To begin with the study of CPSs, first, different types of attacks and their capabilities against physical systems should be investigated. In [17], an attack space defined by the adversary's system knowledge, disclosure, and disruptive resources are introduced, and the attack models are analyzed based on this framework. Also, the effect of each attack on the given plant dynamics is illustrated. Furthermore, a comprehensive classification of cyber-attacks according to the attacker's resources and capabilities can be found in [18] and references therein. To name a few, the Denial of Service (DoS), replay, and covert attacks are well-known attacks that have been studied in the literature [19–21]. The control architecture shown in Figure 1.1 is considered as the CPS reference model. In particular, in Figure 1.1, C-P and P-C denote the communication channels to transfer the signals (control input and sensor measurements) between the plant and the control center. Such an architecture is sufficiently general to include a variety of control system configurations. In such a setup, C-P and P-C channels are considered vulnerable to cyber attacks. In some distributed control systems, the setpoint reference signal is not locally available to the controller but is generated by a distributed command center. In such cases, an attacker can manipulate the setpoint reference signal, preventing the plant from reaching its desired state. This control scenario is commonly seen in domains such as unmanned vehicle formations [22] and smart grids [23]. In other control system applications, the control center is networked, and both control inputs and sensor measurements are transmitted via communication channels. In this setup, these signals may be vulnerable to cyber-attacks. This configuration applies to any control system where the plant is spatially separated from the controller.

From a control perspective, ensuring the security and safety of cyber-physical systems (CPS) requires addressing several key challenges [12]:

- Developing robust attack detection strategies to identify the presence of cyber-attacks
- Designing responsive emergency countermeasures to mitigate the impact of attacks while maintaining system performance at a possibly degraded but acceptable level, such as preserving state and input constraints
- Enabling the control system to recover normal operations and performance levels once

the attack is terminated.

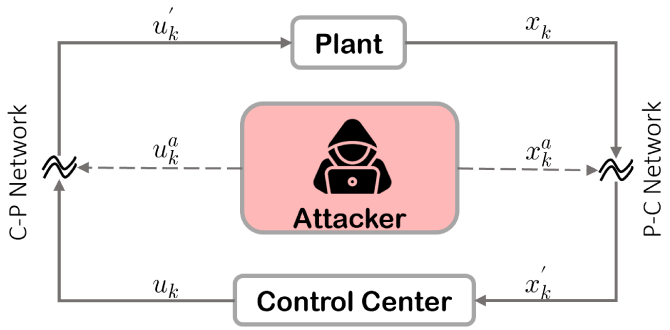


Figure 1.1: Considered Control Architecture

1.2 Cyber-Attacks and Attack Detection Strategies

A significant amount of research has been conducted on the detection of attacks on sensor measurements or control inputs. In [24], a detection mechanism for the detection of covert attacks is proposed by extending system dynamics with an auxiliary system. Another work can be found in [25], where a moving target approach is proposed to detect the presence of stealthy attacks. In addition to moving target approaches, sensor coding approaches have been taken into consideration recently. In [1], a sensor coding solution is presented which capable of detecting stealthy attacks. In this work, stealthy attacks producing infinite estimation errors are investigated, and a sensor encoding mechanism preventing such attacks is derived. For attacks on the sensor measurements, the proposed solution mechanism prescribes the use of a rotation encoding matrix that prevents the injections of the attacks in the stealthy directions related to the unstable eigenvalues of the system. On the other hand, for coordinated attacks on both channels, the encoding strategy prescribes the use of an invertible encoding matrix. In [13], a blended active detection scheme has been proposed that simultaneously uses watermarking and moving target ideas to detect advanced stealthy attacks without any performance degradation. One of the novel approaches for detecting cyber-attacks has been proposed in [26] where by resorting to the concepts of reachability analysis and set membership index, a detection scheme capable of detecting False Data Injection (FDI) and Denial of Service (DoS) attacks

affecting the normal dynamical evolution of the regulated system is presented. In [27], by leveraging reachability arguments and ST-MPC, a networked control architecture is proposed to ensure plant safety in cyber-physical systems (CPS) under cyber-attacks on communication channels. The architecture includes a detector module at the controller to identify false data injection (FDI) attacks on control inputs, sensor measurements, and setpoint signals, and a smart actuator at the plant that triggers an emergency controller to prevent safety violations before they occur.

1.3 Safety and Tracking Performance of CPSs

Another important aspect of CPSs is preserving the safety of the plant and mitigating the effects of cyber attacks on the performance of CPS. Different methods have been proposed to provide safety. In [28], an l_0-based state estimator is proposed to ensure the resilience of the estimation in the presence of attacks. In [29], an observer with an adaptive switching mechanism is proposed to reach an asymptotically stable observation error system under cyber-attacks, and in [30], it is formally proved that a requirement for having a resilient state estimation and control is that the number of under attack sensors should be at most half of the number of available sensors. In [31], a receding-horizon control law is proposed to mitigate the effect of replay attacks. In [32], a mitigation approach against cyber-attacks is proposed by taking advantage of an adaptive resilient control scheme. The problem of designing a resilient control system for a CPS against Denial of Service (DoS) and replay attacks has been studied in [33] and [34], respectively. In [33], using the game theory approach, optimal control strategies have been developed, and based on them, optimality criteria for both defenders and DoS attackers are proposed. In [34], a variation of the receding-horizon control law has been proposed to deal with replay attacks.

One of the novel approaches for attack mitigation is leveraging Set-Theoretic Model Predictive Control (ST-MPC), see e.g., [26, 35]. In [26], by leveraging set-theoretic concepts, watermarking approach, and basic cyber-security tools, a supervisor-based control scheme has been proposed capable of detecting and mitigating FDI and DoS attacks affecting the

normal dynamical evolution of the regulated system. In [35], by leveraging robust reachability arguments, a robust solution to the safety and reference tracking control problems for CPSs has been proposed. The proposed control architecture consists of two add-on modules (one local to the plant and one local to the tracking controller) whose aim is to preserve safety while improving, in a supervised fashion, the tracking performance under attacks.

1.4 Data-Driven Methods in CPS

According to section 1.3, most of the existing methods and architectures for detecting and mitigating cyber attacks on CPSs rely on an accurate model of the system. However, CPSs are often highly complex and have many interacting components, making it difficult to create an accurate mathematical model of the system. Moreover, CPSs are often subject to non-linearities and uncertainties that are difficult to capture in a model-based approach. As a consequence, the use of data-driven methods instead of model-based methods has recently been considered. In contrast to model-based methods, data-driven methods can be used to extract patterns and relationships directly from the system data without requiring a priori knowledge of the underlying physics. Furthermore, data-driven methods can be more robust to these non-linearities and uncertainties because they are able to adapt to changes in the system behavior as they occur. In [36], a data-driven MPC approach has been developed to compute input sequences and predicted outputs obtained from convex programming programs based on pre-collected input-output data. Building on this, a data-driven resilient controller is introduced, ensuring local input-to-state stability under specific denial-of-service (DoS) attacks and noise levels. In [37], a safe learning-based control method through a model predictive safety certification scheme, which enhances potentially unsafe control strategies with safety guarantees and can be combined with any known safe set, has been addressed. By utilizing robust MPC methods, the scheme is suitable for large-scale systems and reduces conservatism over time through a parameter-free, scenario-based design procedure using available and generated closed-loop data. In [38], authors have proposed a modular data-driven safety filter for Cyber-Physical Systems that ensures safety at all times, even during attacks, by validating control commands regardless of their integrity.

This filter operates independently of the control command's reliability and the anomaly detector's characteristics and can be integrated with resilient controllers to maintain high performance during normal operation and safety during attacks.

1.5 Thesis Motivation and Contribution

Based on the reviewed literature in Sections 1.3 and 1.4, there is a gap in research on data-driven detection and mitigation methods for constrained CPS using reachability analysis and set-theoretic control. Hence, in this thesis, two CPS research problems are of interest:

- Design a data-driven detection mechanism for constrained CPSs capable of detecting stealthy and coordinated cyber-attacks.
- Design a data-driven control architecture to preserve the safety and tracking performance of constrained CPSs against cyber-attacks.

1.5.1 Contributions of Thesis

The contributions of the thesis are summarized as follows:

- **Novelty in attack detection mechanism**
    - According to the reviewed literature, some of the existing attack-detection solutions, such as moving target and coding matrix scheme, have the common underlying idea of encoding the sensor measurements and/or control signals before their transmission over the communication channel. However, since the used encoding schemes typically exploit lossless transformations, the encoded data contains (in a different form) the same information available in the not encoded data. As a consequence, if the attacker is able to identify the encoding mechanism, such solutions become ineffective and do not prevent the attacker's possibility of launching undetectable attacks (e.g., replay and covert attacks). In our work [39], we have proposed a novel attack detection strategy based on

a time-varying (random) encoding mechanism, applied to the sensor measurements, based on the concepts of principal components (PCs) analysis, dimensionality reduction, and optimal reconstruction. It has been shown that by using the proposed mechanism, the encoded data transmitted over the network do not have the same information as the original data, i.e., the sensor measurements are compressed in a lower-dimensional latent space. Consequently, the attacker will never have access to the original information. The proposed detection mechanism is capable of detecting stealthy and coordinated cyber attacks such as replay and covert attacks.

- **Novelty in data-driven reachability analysis and ST-MPC**
  - To the best of the author's knowledge, the problem of computing robust Backward Reachable Sets (BRS) (also known as robust one-step controllable sets) from a collection of noisy input-state trajectories and data-driven ST-MPC controllers for constrained linear systems when the model is unknown, has not yet been explored in the literature. Hence, in our work [40], we have developed a novel data-driven method for computing ROSC sets for an unknown linear system subject to bounded disturbances, state, and input constraints. In particular, it has been shown that using a collection of input-output trajectories, a zonotopic inner approximation of the backward reachable sets can be obtained by resorting to an extended space representation. The peculiar feature of the proposed solution is that the computed sets can also be used to efficiently implement a data-driven dual-mode receding horizon controller.

- **Novelty in designing data-driven control architectures capable of preserving the safety and tracking performance of the plant against cyber-attacks**
  - When constrained CPSs are of interest, existing detection and mitigation strategies are developed assuming that an accurate knowledge of the plant's dynamic model is available. As a consequence, recently some progress have been made in the field of data-driven control and safety filters, see, e.g., [41–47]. However,

most of the proposed methods have not been fully exploited to design data-driven anomaly detectors or resilient control strategies for constrained CPSs. Thus, this research work is filling this research gap. In particular, we propose a novel data-driven implementation of the model-based solution presented in [27] for constrained CPSs. To this end, we leverage the data-driven method developed in [48], which is capable of computing zonotopic outer approximations of forward reachable sets to design a novel and robust data-driven passive anomaly detector local to the controller. Moreover, a data-driven set-theoretic predictive controller, which we have developed in [40], is customized to design an emergency controller local to the plant, which is activated whenever the received control input is deemed unsafe according to the expected one-step plant evolution.

- Based on the reviewed literature, most of the existing solutions on the data-driven design of safety preserving control architectures neglect the tracking performance degradation problem under cyber-attacks. In particular, to the best of the author's knowledge, the problem of preserving the safety and tracking performance of a constrained CPS from a collection of noisy input-state trajectories has not yet been explored in the literature. Consequently, this research work goes in the direction of filling the existing gap and developing a solution that tries to minimize, whenever possible and in a data-driven fashion, the tracking performance loss under cyber-attacks. In particular, we here extend the model-based solution in [35] and [2] to develop a novel data-driven control architecture that is capable of preserving the safety of the plant while minimizing, whenever possible, the tracking performance loss due to cyber-attacks on the communication channels.

1.6 Thesis Layout

The main objectives of this research are as follows:

- Design a data-driven detection architecture capable of detecting intelligent and coordinated attacks

- Designing a data-driven control architectures capable of:
    - Detecting cyber-attacks on the communication channels.
    - Preserving plant safety, and minimizing the tracking performance loss under cyber-attacks.

According to these objectives, the manuscript is organized as follows:

- In chapter 2, first, background material on CPSs is provided. Then, some concepts and definitions used along the thesis are presented.
- In chapter 3, a novel active detection mechanism based on the concept of dimentionality reduction and PCA is presented. It is proved that the proposed solution is capable of detecting intelligent and coordinated replay and covert attacks. Moreover, the proposed encoding is equipped with a time-varying random mechanism to prevent the attacker from identifying the PCs of the plant.
- In chapter 4, a data-driven approach for computing backward reachable sets and ST-MPC is introduced. Moreover, to avoid high computational burden in the online phase of ST-MPC, an augmented representation of robust one-step controllable (ROSC) sets is presented and utilized for computing ST-MPC signal. The data-driven strategies and control tools developed in this chapter are instrumental to design the safety preserving architectures in Chapters 5 and 6.
- In chapter 5, a data-driven control architecture which is capable of preserving the safety of constrained CPSs is presented. By leveraging the developed methods in chapter 4, We design three add-on modules—an anomaly detector, a safety verification, and an emergency controller—that enable CPSs to prevent attackers from compromising plant's safety by intervening at least one step before a safety violation occurs.
- In chapter 6, by extending the proposed work in chapter 5, a data-driven control architecture is designed to preserve the safety and also, whenever possible, minimizing

the tracking loss in constrained CPSs. To this end, on the controller side, an add-on module, namely data-driven tracking supervisor, is designed which operates in an open-loop fashion to preserve the tracking performance of the plant in the presence of cyber attacks.

- Finally, in chapter 7, conclusions about the performed research studies are given, and possible future research directions are discussed.

1.7 Publications Related to the Thesis

The following is a list of publications that have resulted from this research:

- [39] **Mehran Attar** and Walter Lucia, "An active detection strategy based on dimensionality reduction for false data injection attacks in cyber-physical systems," *IEEE Transactions on Control of Network Systems*, 2023.

- [40] **Mehran Attar** and Walter Lucia, "Data-driven robust backward reachable sets for set-theoretic model predictive control," *IEEE Control Systems Letters*, 2023.

- [2] **Mehran Attar** and Walter Lucia, "A Data-Driven Safety Preserving Control Architecture for Constrained Cyber-Physical Systems," *International Journal of Robust and Nonlinear Control*, 2024.

- [49] **Mehran Attar**, and Walter Lucia. "A Data-Driven Approach To Preserve Safety and Reference Tracking for Constrained Cyber-Physical Systems Under Network Attacks", *IEEE Transactions on Automatic Control*, 2024. (under review)

Chapter 2

Background, Preliminaries and Definitions

In this chapter, first, background material on CPSs is reviewed. Then, the modus operandi of the dual-mode Set-Theoretic Model Predictive Control (ST-MPC) strategy (used in the successive chapters) is briefly reviewed.

2.1 Networked Control System

In what follows, the main subsystems of the CPS control architecture, shown in Figure 1.1, are presented.

2.1.1 Plant Model

Let us consider the following discrete-time linear time-invariant (LTI) system:

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + w_k \\ y_k &= Cx_k + \eta_k \end{aligned} \tag{2.1}$$

where $k \in \mathbb{Z}_+ := \{0, 1, \ldots\}$, $x_k \in \mathbb{R}^n$ is the state vector, $y_k \in \mathbb{R}^p$ is the output signal, $u_k = f(x_k, r_k) \in \mathbb{R}^m$ is the input vector and w_k and η_k are the process and measurement noise, respectively. Moreover, state and input constraints are prescribed as follows:

$$u(k) \in \mathcal{U}, \quad x(k) \in \mathcal{X} \quad \forall\, k \in \mathbb{Z}_+ \tag{2.2}$$

where $\mathcal{U} \subseteq \mathbb{R}^m$ and $\mathcal{X} \subseteq \mathbb{R}^n$ are compact subsets with $0_m \in \mathcal{U}$ and $0_n \in \mathcal{X}$, respectively. Let us denote with $r_k \in \mathbb{R}^r$ the set-point signal that the output y_k would track.

2.1.2 Control Center

By denoting with $x_k^c \in \mathbb{R}^n$, the state of the controller, its actions can be generically described as

$$u_k = f(x_k^c, y_k, r_k) \tag{2.3}$$

where $f(\cdot, \cdot, \cdot)$, is a stabilizing control logic.

If the C-P and P-C channels are under attack, then (2.1) is changed into:

$$\begin{aligned} x_{k+1} &= Ax_k + B(u_k + u_k^a) + w_k \\ y_k &= Cx_k + y_k^a + \eta_k \end{aligned} \tag{2.4}$$

where u_k^a and y_k^a are the attack vectors on the actuation and measurement channels, respectively.

Definition 2.1 ***(Stealthy Attack)*** *An attack is stealthy if it can reduce the control performance while remaining undetected for an indefinitely large time interval [50].* □

Definition 2.2 ***(Safe Plant Operations)*** *Given a control law $u_k := f(x_k, r_k)$. The plant (2.1) closed-loop evolution under $f(x_k, r_k)$ is considered safe if the prescribed constraints (6.4) are satisfied.* □

Definition 2.3 ***(Attack with disclosure and disruptive resources)*** *Let us consider an insecure communication channel, namely channel $-\, i$, where the data packet $d \in \mathbb{R}^v$ is transmitted at each sampling time $t \in \mathbb{Z}_+$.*

- *Disclosure Resource: An attacker has disclosure resource on the channel $-\, i$ if he/she can violate the confidentiality property, i.e., read the vector $z(t)$ for an arbitrarily large time interval.*
- *Disruptive Resource: An attacker has disruptive resources on the channel $-\, i$ if he/she can violate the authentication or integrity property, i.e., the attacker can arbitrarily alter the transmitted vector z into a new compatible vector $z' \in \mathbb{R}^z$.* □

Definition 2.4 *[51] (**False Data Injection Attack**) Let's consider two subsystems, S_1 and S_2, and the communication channel between them. By denoting with $v \in \mathbb{R}^v$ the vector transmitted from S_1 and with $v' \in \mathbb{R}^v$ the signal received by S_2, we define a False Data Injection (FDI) attack on v, a network attack capable of changing v_k by adding an arbitrary vector v_k^a,, i.e.,*

$$v_k^{'} = v_k + v_k^a \tag{2.5}$$

Definition 2.5 *(**Attack with full model knowledge**) An attack with full-model knowledge is an attack with perfect knowledge about the whole closed-loop system behaviors and dynamics (system plant, controller, detector).*

2.2 Cyber-Attack Classification

In this section, different attack scenarios are classified according to their model knowledge and resources. Since the studied attacks in the literature are all on the C-P and P-C channels, and the only studied attack on the C-C channel is FDI attack, we only introduce the first group. To classify cyber-attacks, a 3-D attack space model which categorizes attacks based on their *a-priori system model knowledge*, *disclosure* and *disruptive resources* is provided in [17] (see Fig. 2.1). In the following, these attacks space model is presented, and the available information and resources for each attack are specified.

2.3 Attacker's Knowledge and Resources

When the security of communication channels is of interest, then the main properties are of interest: *Confidentiality*, *Integrity* and *Availability* (also known as the CIA triad) [52]. Violating any of those properties causes security and privacy issues. The definition of each property in CPS is as follows:

- **Confidentiality:** Refers to the ability to keep information secret from unauthorized users. If the adversary could not obtain/read information/data related to the system by eavesdropping on the communication channels, the NCS has confidentiality.

- **Integrity:** Refers to the trustworthiness of data or resources. If the adversary could not alter transmitted information/data on the communication channels, the NCS has Integrity.

- **Availability:** Refers to the ability of a system/data to be accessible and usable upon demand. If the adversary cannot prevent the data to be reached to the receiver, the NCS has availability.

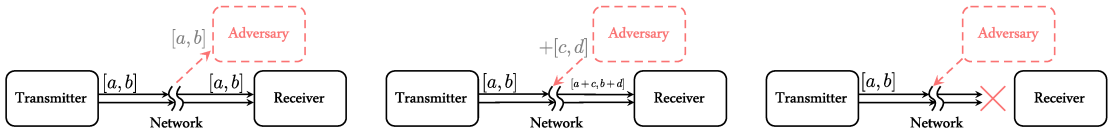


Figure 2.1: Lack of (2.1a) Confidentiality. (2.1b) Integrity (2.1c) Availability

Attacks on CPSs could be categorized based on the set of resources available to the attacker. The more resources the attacker has, the more complex attacks can be performed. As an example, Fig. 2.2 shows the required resources and level of system knowledge for performing different well-studied types of attacks.

- **Disclosure resources:** is the set of communication channels where the attacker can violate the confidentiality property (read/obtain data/information from the networked communication channel)

- **Disruptive resources:** is the set of communication channels where the attacker can violate the integrity and/or availability property (alter data/information on the networked communication channel)

- **System model knowledge:** the adversary's knowledge and information about the plant, controller, and any other used dynamics in the system.

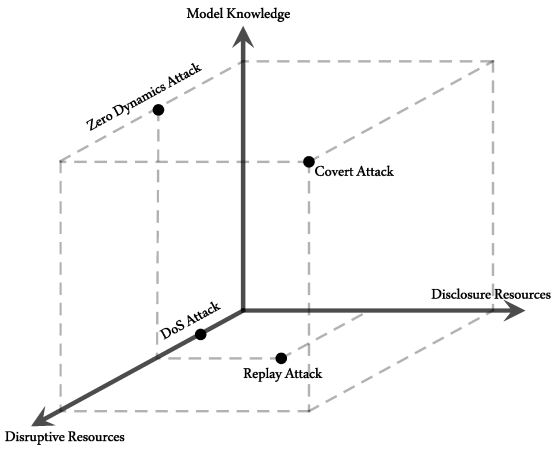


Figure 2.2: Cyber-Physical Attack Space

2.4 Different Types of Attacks

In this section, different cyber-attack scenarios are explained.

2.4.1 Denial of Service (DoS) Attack

In this attack scenario, the adversary tries to prevent transmitted data from reaching their destination. Therefore, the receiver will not be able to receive the signal properly, and the whole system functionality will be affected. Due to its behavior, the DoS attack could be misrecognized with a poor network connection. Performing a DoS attack does not need any system model knowledge and disclosure resources (as shown in Fig. 2.2). Usually, this attack is performed by jamming the communication channels using disruptive resources.

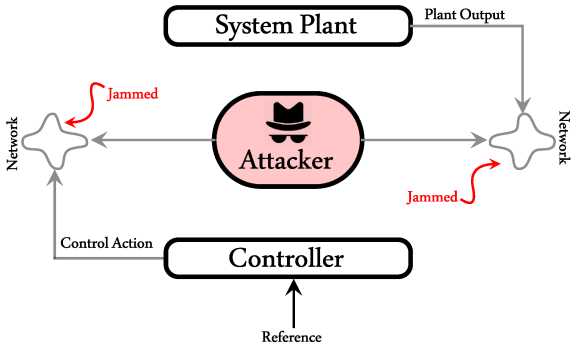
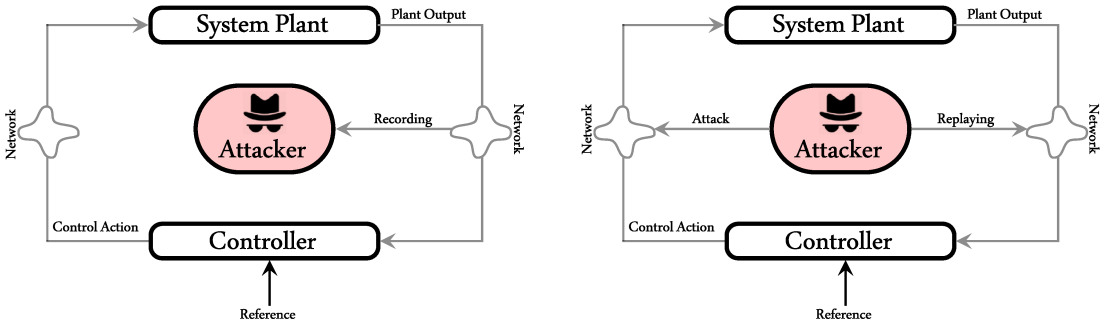


Figure 2.3: DoS Attack on CPS

2.4.2 Replay Attack

In this attack scenario, the adversary tries to replay valid previously recorded measurement data y_k (plant output) while injecting attacked signals in the actuation channel. Therefore, the controller receives valid but fake measurement signals from the plant. Performing a replay attack needs disclosure resources on the measurement channel and disruptive resources on both channels. This attack scenario is performed in two phases (see Fig. 2.4):

- **Phase I. Recording:** The adversary records the measurement signals for a period of time.
- **Phase II. Replaying:** The recorded data are replayed while malicious inputs are injected into the actuation channel.



(a) (b)

Figure 2.4: Replay attack (2.4a) phase-1. (2.4b) phase-2

Remark 2.1 *The replay attack is an FDI attack with disruptive resources on both communication channels and disclosure capabilities on the measurement channel. A replay attack is performed as follows: in the actuation channel, an arbitrary vector u_k^a is injected to deteriorate the control system performance, while, in the measurement channel, previously recorded measurements are used to replace the actual sensor measurements. A replay attack is stealthy if performed in steady-state conditions [20].* □

2.4.3 Covert Attack

Consider the plant model (2.1). The covert attack consists of injecting an arbitrary FDI attack on the actuation channel whose effect on the plant dynamics is properly canceled out from the sensor measurement to avoid detection. In particular, first, the attacker injects an FDI control input attack u_k^a into the C-P channel:

$$u'_k = u_k + u_k^a \tag{2.6}$$

Then, the effect of this attacked control input on the plant dynamics is removed from the sensor measurement vector by performing the following FDI attack:

$$y'_k = y_k - y_k^a \tag{2.7}$$

where y_k^a is the effect of attack u_k^a on the plant dynamics, such that:

$$y_k^a := -C \sum_{j=0}^{k-1} A^j B u_{k-1-j}^a \tag{2.8}$$

Performing covert attacks requires full model knowledge about the plant dynamics and disruptive resources on both actuation and measurement channels.

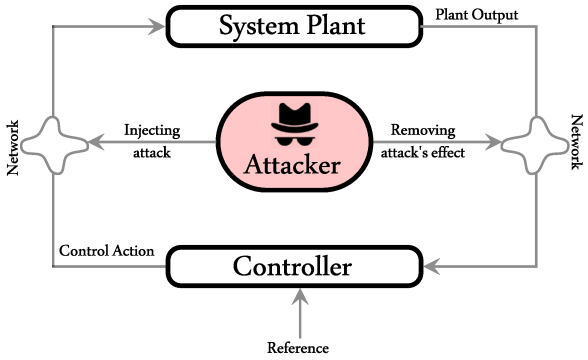


Figure 2.5: Covert Attack

Remark 2.2 *The covert attack introduced is an FDI attack with disclosure and disruptive resources on both communication channels and model knowledge about the plant dynamical model (2.1). The covert attack actions are the following: in the actuation channel,*

an arbitrary vector u_k^a is added to deteriorate the control system performance, while, in the measurement channel, a properly designed vector y_k^a is injected to remove the input attack effect from the measurements. A covert attack is stealthy regardless of the detection rule/mechanism used in the control center [21]. □

2.5 Detection Mechanisms

In the fault diagnosis literature, several methods have been proposed to deal with malfunctions in control systems [53], [54], and the same strategies can also be used to detect anomalies caused by cyber-attacks. DoS attacks, because of their nature, can be easily detected even with passive detectors. However, it has been proved that such detectors have limitations, and some specialized attacks such as covert, replay, and zero-dynamics attacks cannot be detected by them [6], [55]. Therefore, active detection mechanisms are proposed to deal with these types of attacks. Active defensive strategies are based on limiting disruptive and disclosure resources available to the attacker or adding information to the plant that is unknown to the attacker. In what follows, different detection mechanisms that have been proposed to deal with the aforementioned well-known attacks are presented.

2.5.1 Watermarking Approach

The watermarking-based detection mechanism was first proposed in [56] and is mostly used for the detection of replay attacks on the measurement channel. The main idea of this active mechanism is to inject private excitation into the system to reveal malicious behavior.

$$u_k = u_k^* + \Delta u_k \tag{2.9}$$

Where u_k^* is the calculated optimal control input, and Δu_k is a zero-mean IID Gaussian distribution that acts as an authentication signal.

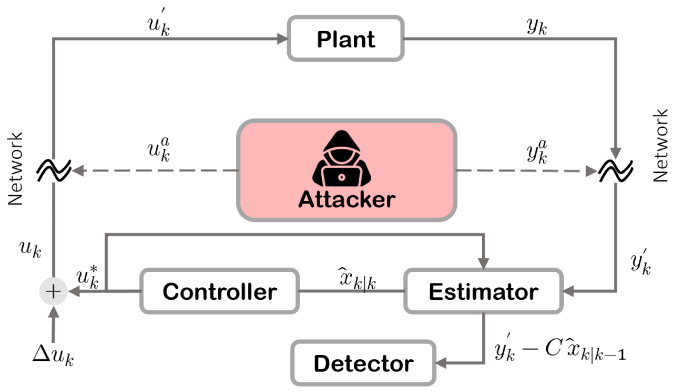


Figure 2.6: Watermarking Approach

2.5.2 Sensor Coding Approach

The sensor coding detection mechanism is presented in [1]. As its name shows, it includes encrypting the sensor measurement data in a way that the adversary could not obtain meaningful information about the sensors and even the plant dynamics. Proper sensor coding could reveal stealthy false data injections like covert attacks.

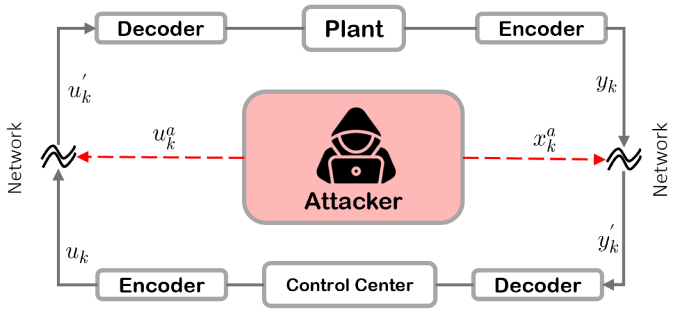


Figure 2.7: Sensor Coding Approach

2.5.3 Moving Target Approach

The moving target detection mechanism is presented in [24], [25] to detect covert and zero-dynamics attacks. It prescribes coupling the system plant with an auxiliary system or dynamics unknown to the attacker. The latter enables a passive detector on the control side to detect attacks.

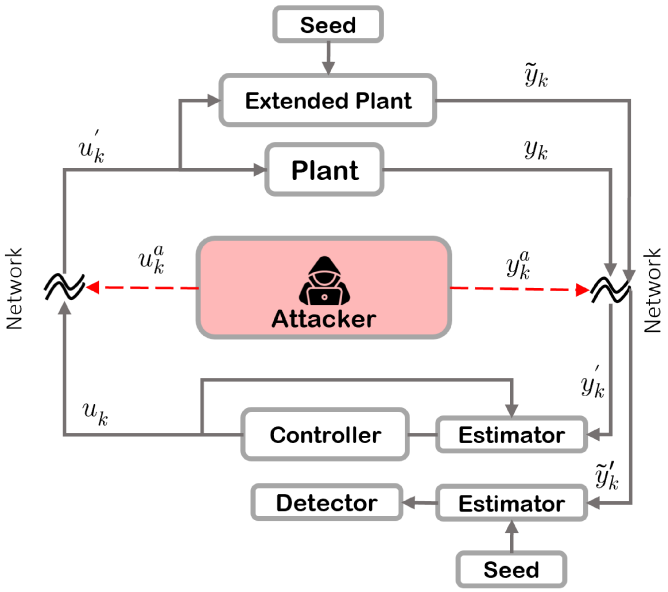


Figure 2.8: Moving target approach

2.6 Safety

Ensuring the safety of the physical plant in the presence of cyber-attacks is of paramount importance. This can be described in Fig. 2.9, where the regions of required, degraded, and unacceptable performances are shown for a generic CPS. In general, the attacker aims to force the physical plant state trajectory to move from the region of required performance toward the region of unacceptable performance while the defender tries to keep the system within the region of required performance. A CPS is said to be safe in the presence of attacks in the region of unacceptable performance is never reached.

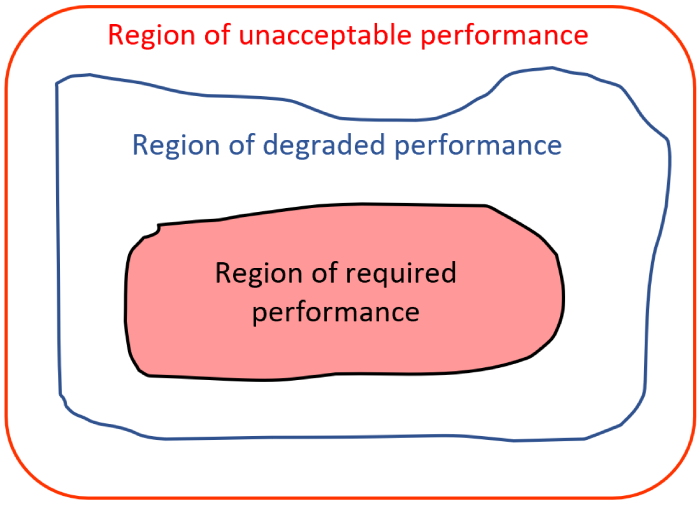


Figure 2.9: Working Configurations

2.7 Set-Theoretic Model Predictive Control

In this section, the dual-mode MPC controller developed in [57], hereafter denoted as ST-MPC, is summarized. Consider the linear system described in (2.1). Moreover, assuming that due to physical limitations and safety reasons, the following set-membership constraints are prescribed:

$$x_k \in \mathcal{X} \subset \mathbb{R}^n, \quad u_k \in \mathcal{U} \subset \mathbb{R}^m, \tag{2.10}$$

where we assume that $\mathcal{X} \subset \mathbb{R}^n, \mathcal{U} \subset \mathbb{R}^m, \mathcal{W} \subset \mathbb{R}^n$ are compact sets. The following definitions are instrumental for computing ST-MPC controller.

Definition 2.6 ***Robust Control Invariant (RCI)*** *A set $\mathcal{T}^0 \subseteq \mathcal{X}$ is called Robust Control Invariant (RCI) for* (2.1)-(2.10) *if $\forall x \in \mathcal{T}^0, \exists u \in \mathcal{U} : Ax + Bu + w \in \mathcal{T}^0, \forall w \in \mathcal{W}$ [58, Definition 11.22].*

Definition 2.7 ***Robust One-Step Controllable (ROSC)*** *Consider* (2.1)-(2.10) *and a target set $\mathcal{T}^j \subseteq \mathcal{X}$. The set of states Robust One-Step Controllable (ROSC) to $\mathcal{T}^j$ is*

$$\mathcal{T}^{j+1} = \{x \in \mathcal{X} : \exists u \in \mathcal{U} : Ax + Bu + w \in \mathcal{T}^j, \forall w \in \mathcal{W}\} \tag{2.11}$$

A stabilizing receding-horizon controller capable of fulfilling all the prescribed constraints and capable of robustly confining, in a finite number of steps, the state trajectory into terminal RCI set $\mathcal{T}_0$ can be designed as follows [26, Section II.A]:

- *Offline:*

(1) Consider the constraint-free and disturbance-free model (2.1) and compute a stabilizing state-feedback control law, e.g., $u_k = -Kx_k$, where $(A - BK)$ is a stable matrix. Then, compute the smallest RCI region associate to $\mathcal{T}^0$ for (2.1) with (2.10), see, e.g., [59].

(2) Starting from $\mathcal{T}^0$, recursively compute a sequence of $N > 0$ ROSC sets $\{\mathcal{T}^j\}_{j=1}^{N}$, where

$$\mathcal{T}^j = \{x \in \mathcal{X} : \exists u \in \mathcal{U} : x^+ \in \mathcal{T}^{j-1}, \forall w \in \mathcal{W}\} \tag{2.12}$$

- *Online ($\forall\, k$):*

(1) Find $j_k := \min_{j \in \{0,\ldots,N\}} \{j : x_k \in \mathcal{T}^j\}$

(2) **If** $j_k = 0$, **then** $u_k = -Kx_k$. **Else** solve the following Quadratic Programming (QP) problem:

$$\begin{aligned} u_k = \arg\min_{u \in \mathcal{U}} J(x_k, u) \text{ s.t.} \\ Ax + Bu \in (\mathcal{T}^{j_k - 1} \ominus \mathcal{W}) \end{aligned} \tag{2.13}$$

where $J(x_k, u)$ is a convex cost function.

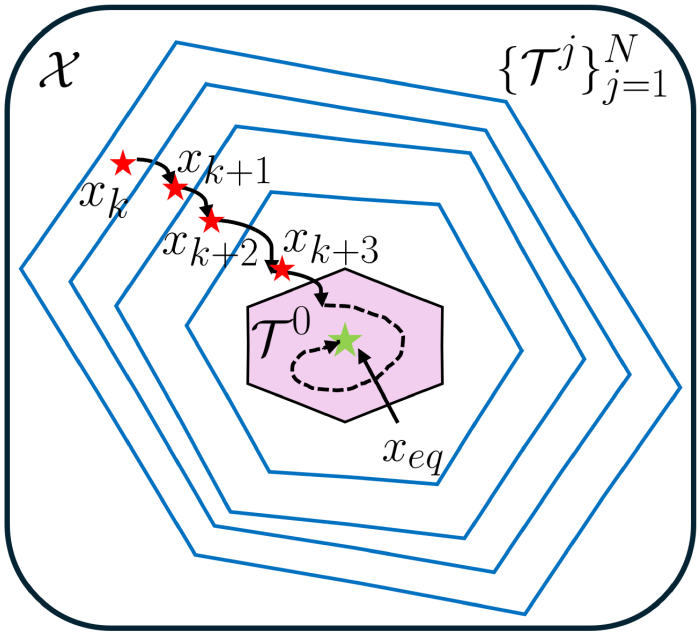


Figure 2.10: Family of robust one-step controllable sets, and the state trajectory evolution

Remark 2.3 *Please note that the above one-step controllable sets can be numerically computed in Matlab, using, e.g., MPT3 toolbox [60].*

Property 2.1 *The ST-MPC algorithm enjoys the following properties [57]:*

(1) The optimization (2.13) *enjoys recursive feasibility.*

(2) The terminal region $\mathcal{T}^0$ is reached in at most j_0 steps (with j_0 the set membership index at $k = 0$) regardless of any admissible disturbance realization.

2.8 Set Representation and Set Operations

The following definitions are taken or adapted from [48, 58, 61].

Definition 2.8 *Given two sets $\mathcal{S}_1$ and $\mathcal{S}_2$, the Minkowski/Pontryagin set sum (denoted as $\oplus$) and difference (denoted as $\ominus$) between $\mathcal{S}_1$ and $\mathcal{S}_2$ are:*

$$\begin{aligned} \mathcal{S}_1 \oplus \mathcal{S}_2 &= \{s_1 + s_2 : s_1 \in \mathcal{S}_1, s_2 \in \mathcal{S}_2\} \\ \mathcal{S}_1 \ominus \mathcal{S}_2 &= \{s_1 \in \mathbb{R}^s : s_1 + s_2 \in \mathcal{S}_1, \forall s_2 \in \mathcal{S}_2\} \end{aligned} \tag{2.14}$$

Definition 2.9 ***Polytope*** *A polytope $\mathcal{P}$ in $\mathbb{R}^n$ is defined as the intersection of a finite set of q closed halfspaces in $\mathbb{R}^n$:*

$$\mathcal{P} = \left\{ x \in \mathbb{R}^n \,|Cx \leq d, C \in \mathbb{R}^{q \times n}, d \in \mathbb{R}^{q \times 1} \right\} \tag{2.15}$$

The definition of $\mathcal{P}$ as in (2.15) is known in the literature as the Half-space representation ($\mathcal{H}$-representation) of $\mathcal{P}$.

Definition 2.10 ***Zonotope*** *Given a center vector $c \in \mathbb{R}^n$ and $p \in \mathbb{N}$ generator vectors $g^{(i)} \in \mathbb{R}^n$ collected in a matrix $G = \left[g^{(1)} \ \ldots, \ g^{(p)} \right] \in \mathbb{R}^{n \times p}$, referred to as the generator matrix. Then, a zonotope is defined as*

$$\mathcal{Z}(c, G) = \left\{ x \in \mathbb{R}^n : x = c + \sum_{i=1}^{p} \beta^{(i)} g^{(i)}, -1 \leq \beta^{(i)} \leq 1 \right\} \tag{2.16}$$

The definition of $\mathcal{Z}(c, G)$ as in (2.16) is known in the literature as the Generator representation ($\mathcal{G}$-representation) of $\mathcal{Z}(c, G)$ (see Figure 2.11).

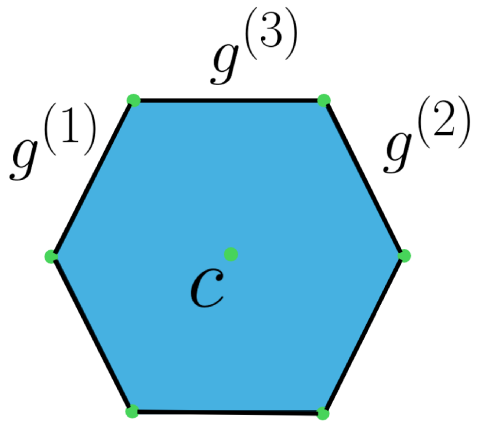


Figure 2.11: Zonotope

Definition 2.11 ***Matrix Zonotope*** *Given a center matrix $C \in \mathbb{R}^{n \times p}$ and $q \in \mathbb{N}$ generator matrices $G_M^{(i)} \in \mathbb{R}^{n \times p}$ collected in a matrix $G_M = \left[G_M^{(1)}, \ldots, G_M^{(q)} \right] \in \mathbb{R}^{n \times (pq)}$. Then, a matrix zonotope is*

$$\mathcal{M}(C, G_M) = \{ X \in \mathbb{R}^{n \times p} : X = C + \sum_{i=1}^{q} \beta^{(i)} G_M^{(i)}, \quad -1 \leq \beta^{(i)} \leq 1 \} \tag{2.17}$$

Definition 2.12 ***Matrix Polytope*** *Consider a set of $n_v > 0$ vertex matrices $\mathcal{V}_P = \{V_P^{(i)}\}_{i=1}^{n_v}$, $V_P^{(i)} \in \mathbb{R}^{n \times p}$. A matrix polytope $\mathcal{M}_P(\mathcal{V}_P) = \{M \in \mathbb{R}^{n \times p} : M = \sum_{i=1}^{n_v} \rho_i V_P^{(i)}, 0 \leq \rho_i \leq 1, \sum_{i=1}^{n_v} \rho_i = 1\}$.*

Definition 2.13 *Let $\mathcal{M}_1, \mathcal{M}_2$ be two matrix sets (e.g., matrix zonotopes) and $\mathcal{Z}$ a vector set (e.g., a zonotope). Then,*

$$\begin{aligned} \mathcal{M}_1 \mathcal{M}_2 &= \{M_1 M_2 : M_1 \in \mathcal{M}_1,\ M_2 \in \mathcal{M}_2\} \\ \mathcal{M}_1 \mathcal{Z} &= \{M_1 z : M_1 \in \mathcal{M}_1,\ z \in \mathcal{Z}\} \end{aligned} \tag{2.18}$$

Chapter 3

An Active Detection Strategy Based on Dimensionality Reduction for False Data Injection Attacks in Cyber-Physical Systems

3.1 Introduction

In recent years, different intelligent and undetectable cyber-attacks against networked control systems (e.g., replay and covert attacks) have been investigated. In this chapter, we design a novel control architecture that prevents their existence. In particular, first, we propose to encode the sensor outputs into a randomly changing lower-dimensional space obtained by means of principal component analysis. Such a transformation ensures optimal information reconstruction on the controller's side, and it prevents the attacker from accessing the original sensor measurements, nullifying the possibility of perfect stealthy attacks. Then, on the controller's side, we design a passive Gaussian anomaly detector that leverages the output of an unknown input observer ad-hoc designed to estimate the system's states and unknown inputs simultaneously. It is formally shown that the proposed detection strategy is able to discover the presence of replay and covert attacks. Simulation results obtained considering a quadruple water tank system confirm the capability of the developed architecture.

3.1.1 Contribution of The Work

According to the literature, some of the existing attack-detection solutions, such as moving target and coding matrix scheme, have the common underlying idea of encoding the sensor measurements and/or control signals before their transmission over the communication channel. However, since the used encoding schemes typically exploit lossless transformations, the encoded data contains (in a different form) the same information available in the not encoded data. As a consequence, if the attacker is able to identify the encoding mechanism, such solutions become ineffective and do not prevent the attacker's possibility of launching undetectable attacks (e.g., replay and covert attacks).

We here propose a novel attack detection strategy based on a time-varying (random) encoding mechanism, applied to the sensor measurements, based on the concepts of *principal components* (PCs) analysis, dimensionality reduction, and optimal reconstruction. The main novel features of the proposed solution can be summarized as follows: (i) the encoded data, transmitted over the network, do not have the same information as the original data, i.e., the sensor measurements are compressed in a lower-dimensional latent space. Consequently, the stealthy attack directions that might be present in the actual sensor measurements are not usable on the lower-dimensional transmitted data. This is different from the solution in [1], where the coding matrices must be ad-hoc designed to rotate the stealthy attack directions. Moreover, unlike [1], the proposed encoding mechanism is derived without any assumptions about the stability of the system, and the attack detection strategy is not limited to attacks producing an unbounded state estimation error; (ii) the proposed solution ensures that FDI attacks on the sensor measurements, replay, and covert attacks are not stealthy. Moreover, the proposed randomly changing encoding mechanism cannot be identified by an attacker eavesdropping on the communication channels; (iii) the information loss associated with the encoding scheme is minimal (among all possible linear mappings), bounded and a-priori known. Moreover, to deal with the reconstruction loss, we model the introduced error as an unknown disturbance that affects the sensor outputs of the system. By doing so, we then develop an *Unknown Input Observer* (UIO) [62–66], based on the UIO scheme developed in [67]. The proposed UIO can estimate the disturbances

generated by the dimensionality reduction and, differently from [67], it can deal with False Data Injections (FDIs) that can change over time and are not sparse. Finally, by leveraging the time-varying encoding mechanism and UIO, we develop an active data-driven Gaussian anomaly detector capable of ensuring the absence of stealthy attacks.

Please note that the problems of ensuring the stability and performance of the closed-loop system under attacks are not addressed in this work. Nevertheless, the here designed UIO can be used to develop proper control mitigation strategies, similar to the solutions proposed in [68, 69]. Such a control problem is left for future works.

3.2 Problem Formulation

Consider a networked control system (see Fig. 3.1) where both actuation and measurement channels are prone to cyber-attacks.

3.2.1 Plant Model

We consider the class of discrete-time Linear Time-Invariant (LTI) systems, modelled by

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + w_k \\ y_k &= Cx_k + \eta_k \end{aligned} \tag{3.1}$$

where $x_k \in \mathbb{R}^n$, $u_k \in \mathbb{R}^{m_1}$ and $y_k \in \mathbb{R}^p$ are state, input and output measurement vectors, respectively, while, A, B, C are matrices with proper dimensions. We assume that the pair (A, C) is detectable. Moreover, $w_k \in \mathbb{R}^n$, $\eta_k \in \mathbb{R}^p$ are truncated zero-mean Gaussian with a-priori unknown covariance matrix which model the process noise and sensor measurement noise, respectively. A stabilizing state-feedback controller is assumed to be already available in the control center, while the anomaly detector and state estimator will be designed.

3.2.2 Networked Cyber-Attacks: Models and Resources

We assume that the attacker is aware of the plant's model (3.1) and that it is capable of corrupting the integrity and confidentiality of the sensor measurements y_k and control signal u_k once they are transmitted over the communication channel. FDI attacks on the

communication channels are modeled as follows:

$$u'_k := u_k + u^a_k, \quad y'_k := y_k + y^a_k \tag{3.2}$$

where $u^a_k \in \mathbb{R}^{m_1}$ and $y^a_k \in \mathbb{R}^p$ are possible unbounded attack vectors.

Problem statement: Of interest in this chapter is the problem of designing a novel active detection mechanism based on the concept of dimensionality reduction, capable of ensuring the absence of stealthy replay and covert attacks.

3.3 Proposed Architecture for Attack Detection

In this section, first, we introduce an encoding scheme based on the dimensionality reduction concept to prevent the attacker from accessing the original sensor measurements. Then, we analyze its impacts on the decoded sensor measurements (on the controller's side of the CPS), showing that the introduced error can be modeled as a bounded additive disturbance on the system's outputs. Therefore, we design a stable UIO to estimate the combined effect of measurement disturbances produced by the sensor's noise and encoding scheme. Finally, we leverage a time-varying encoding scheme (acting as a moving target) and the UIO to design a data-driven Gaussian-based anomaly detector to reveal FDI attacks and ensure the absence of stealthy attacks.

The standard networked control system scheme is modified to prevent the attacker from accessing the original sensor measurements by representing sensor outputs in a lower-dimensional space. In particular, here, this is done by adding an encoder and decoder blocks to the plant-side and control center, respectively (see Fig. 3.1). The encoder block encodes, on the plant's side, the sensor outputs in a lower-dimensional space using the principal components of sensor outputs. In the control center, the decoder block reconstructs the transformed data in the original space. In what follows, the concept of optimal dimensionality reduction is presented.

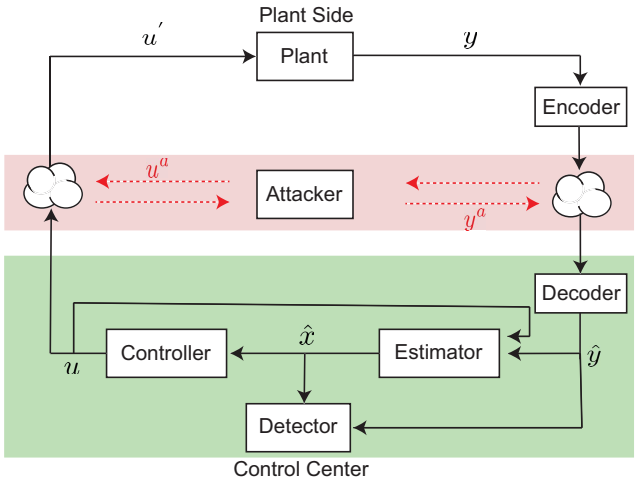


Figure 3.1: Proposed networked control system architecture.

3.3.1 Dimensionality Reduction and Principal Components

Consider the following linear mapping

$$\tilde{y}_k = \phi y_k, \quad \phi \in \mathbb{R}^{p \times p} \tag{3.3}$$

where ϕ is an orthonormal matrix. Of interest is the transformation matrix ϕ^* minimizing the reconstruction error $\|y_k - \hat{y}_k\|^2$ where $\hat{y}_k = \phi^{*T}\phi^* y_k$, i.e.,

$$\begin{aligned} \phi^* &= \arg\min_{\phi} \|y_k - \hat{y}_k\|_2^2 \\ &\text{s.t.} \quad \phi\phi^T = I_p \end{aligned} \tag{3.4}$$

To solve the above optimization problem, we use the concept of principal components. Consider $\Delta k = \left[k_1, k_2\right]$ as a time period in which the system's evolution is in the attack-free scenario, and k_1 is sufficiently large to ensure that the state estimation error has converged to zero. By collecting N samples from each sensor output during Δk, we can define the matrix $Y^{(N)} \in \mathbb{R}^{p \times N}$ and its normalization, namely $\bar{Y}^{(N)}$ (see the notation subsection). Then, the optimization (3.4) can be defined as follows:

$$\begin{aligned} &\min_{\phi} \frac{1}{2N} \sum_{j=1}^{N} \|\bar{Y}_{*j}^{(N)} - \hat{\bar{Y}}_{*j}^{(N)}\|_2^2 \\ &\text{s.t.} \quad \phi\phi^T = I_p \end{aligned} \tag{3.5}$$

where $\hat{\bar{Y}}^{(N)}_{*j} = \phi^T \phi \bar{Y}^{(N)}_{*j}$, $\forall\, j$, are the reconstructed samples. It can be proved that the solution to the above optimization problem is equivalent to finding the directions that maximize the variance of $\bar{Y}^{(N)}$ [70]. Thus, we can redefine this optimization problem as follows:

$$\begin{aligned} \max_{\phi} \quad & \phi S \phi^T \\ \text{s.t.} \quad & \phi \phi^T = I_p \end{aligned} \tag{3.6}$$

where $S = (\bar{Y}^{(N)})(\bar{Y}^{(N)})^T$. By using a diagonal Lagrange multiplier matrix $\Lambda \in \mathbb{R}^{p \times p}$, the Lagrangian of (3.6) is

$$L(\phi, \Lambda) = \phi S \phi^T - \Lambda(\phi \phi^T - I_p) \tag{3.7}$$

By imposing $\frac{\partial L}{\partial \phi} = 0$, we have

$$S\phi = \Lambda \phi \tag{3.8}$$

where Λ contains the eigenvalues λ_j of S, and each column of ϕ i.e., ϕ_{*j}, is an eigenvector of S associated to λ_j. If we re-order λ_j in Λ and ϕ_{*j} in S such that $\lambda_1 \geq \ldots \geq \lambda_p$, then ϕ_{*1} is known as the direction of maximum variation [71]. Similarly, $\phi_{*2}, \ldots, \phi_{*p}$ are, in a descending order, the next directions of the maximum variation. In the literature, ϕ_{*j} are known as the Principal Components (PC) of $\bar{Y}^{(N)}$ [71]. Given ϕ, it is possible to define the following set of linear transformations, which map vectors of $\mathbb{R}^p$ into a lower-dimensional space $\mathbb{R}^d$, with $d \leq p$.

$$\mathcal{T} = \{ \underbrace{\left[\phi_{*1}\right]}_{\Phi_1}, \underbrace{\left[\phi_{*1}|\phi_{*2}\right]}_{\Phi_2}, \ldots, \underbrace{\left[\phi_{*1}|\phi_{*2}|, \ldots, |\phi_{*p}\right]}_{\Phi_p} \} \tag{3.9}$$

Therefore, by applying the semi-orthogonal matrix Φ_d to the sensor measurements y_d, i.e.,

$$\tilde{y}_k = \Phi_d^T y_k \tag{3.10}$$

we project $y_k \in \mathbb{R}^p$ into a lower-dimensional vector $\tilde{y}_k \in \mathbb{R}^d$. Moreover, the decoder

$$\hat{y}_k = \Phi_d \tilde{y}_k \in \mathbb{R}^p \tag{3.11}$$

minimizes the reconstruction error $\|y_k - \hat{y}_k\|_2$.

Remark 3.1 *The operation* (3.10) *projects the sensor measurements along with the directions of maximum variation. Consequently, if the sensor outputs contain independent identically distributed Gaussian noise, the projected sensor outputs (in a lower-dimensional space) will contain a lower amount of noise [72]. In other words, the projection along the principal components is beneficial to decrease the sensor noise.*

Remark 3.2 *Suppose the attacker is unaware of the projection matrices* Φ_d *in* (3.10)*. In that case, the attacker won't be able to either reconstruct the original sensor measurement vectors or perform a system identification of the plant's dynamics. The following sections will address the case where the set* $\mathcal{T}$ *is known to the attacker.*

3.3.2 Reconstruction Error Bound

In this section, we analyze the upper bound of the reconstruction error $\|e_k^{dr}\|_2 = \|y_k - \hat{y}_k\|_2$ produced by the encoder (3.10) and decoder (3.11) for $d < p$.

If at the time instant k, the transformation matrix Φ_r is applied to y_k, then the reconstruction error vector e_k is:

$$e_k^{dr} = \hat{y}_k - y_k = \Phi_r \Phi_r^T y_k - y_k = \mathcal{M} y_k \tag{3.12}$$

where $\mathcal{M} = \Phi_r \Phi_r^T - I_p$. Therefore, by considering the Euclidean norm of the reconstruction error, it is possible to write that

$$||e_k^{dr}||_2 = ||\mathcal{M} y_k|| \leq \rho(\mathcal{M}) ||y_k||_2 \tag{3.13}$$

where $\rho(\mathcal{M}) = \lambda_{max}(\mathcal{M})$ is the largest eigenvalue of $\mathcal{M}$. As a consequence, the reconstruction error e_k^{dr} can be considered an unknown bounded output disturbance acting on the sensor measurements y_k.

3.4 FDI Estimation Using a UIO

Given the bounded reconstruction error e_k^{dr} introduced by the encoder/decoder operations, the plant's measurement vector received by the state estimator/controller can be written as

$$\hat{y}_k = Cx_k + \eta_k + e_k^{dr} = Cx_k + v_k \tag{3.14}$$

where $v_k = \eta_k + e_k^{dr}$. Therefore, the following plant model can be considered for estimation purposes:

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + Eq_k \\ \hat{y}_k &= Cx_k + Dq_k \end{aligned} \tag{3.15}$$

where $q_k = \begin{bmatrix} w_k^T & v_k^T \end{bmatrix}^T$, $E = \begin{bmatrix} I_n & 0_n \end{bmatrix}$ and $D = \begin{bmatrix} 0_p & I_p \end{bmatrix}$. As a consequence, for (3.15) it is possible to design a UIO (see Fig. 3.2) to simultaneously estimate x_k and the non-sparse unknown inputs/output disturbances w_k and v_k.

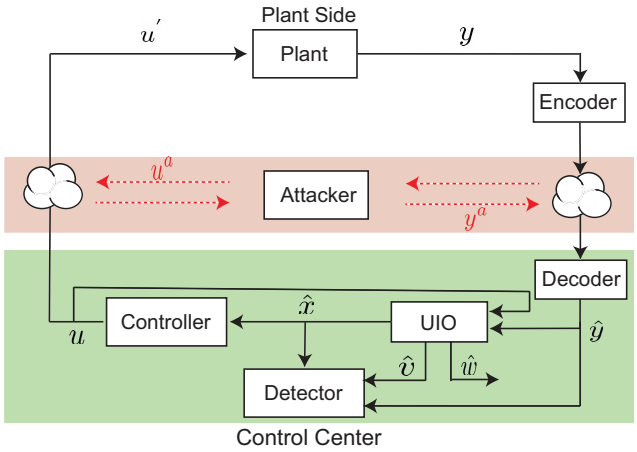


Figure 3.2: Proposed networked control system architecture with UIO.

Consider the following full-state UIO architecture [73]

$$\begin{aligned} z_{k+1} &= Nz_k + Gu_k + L\hat{y}_k \\ \hat{x}_k &= z_k + M\hat{y}_k \end{aligned} \tag{3.16}$$

and let $e_k = x_k - \hat{x}_k$ be the UIO state estimation error. Then, the dynamic evolution

of e_k is

$$\begin{aligned} e_{k+1} &= x_{k+1} - \hat{x}_{k+1} \\ &= Ne_k + \mathcal{P}_1 x_k + \mathcal{P}_2 u_k + \mathcal{P}_3 q_k - \mathcal{P}_4 q_{k+1} \end{aligned} \tag{3.17}$$

where $\mathcal{P}_1 = A+NMC-N-LC-MCA$, $\mathcal{P}_2 = B-G-MCB$, $\mathcal{P}_3 = E+NMD-LD-MCE$ and $\mathcal{P}_4 = MD$. The error dynamic (3.17) is asymptotically stable if and only if the following conditions hold [73]:

- N is asymptotically stable
- $\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3, \mathcal{P}_4 = 0$

Since D is full row rank, then $\mathcal{P}_4$ is a zero matrix only if $M = 0$. Moreover, if $M = 0$, the UIO has the structure of Luenberger-like observer and the dynamic of the estimation error is:

$$e_{k+1} = Ne_k + [A - N - LC]x_k + [B - G]u_k + (E - LD)q_k \tag{3.18}$$

By imposing $\mathcal{P}_1 = 0$, $\mathcal{P}_2 = 0$, then $N = A - LC$, $B = G$, and (3.18) becomes

$$e_{k+1} = (A - LC)e_k + (E - LD)q_k \tag{3.19}$$

Then, if there exists L such that $(A - LC)$ is Schur stable and $(E - LD) = 0_{n\times(n+p)}$, then the error dynamic (3.19) is asymptotically stable. However, in general, if this is not possible, the resulting error is bounded, see e.g. [74]. The design of the gain L and disturbance q_k estimation is addressed in the following sections.

3.4.1 Error Dynamics Stability

In this subsection the estimator gain L is designed to ensure globally uniformly l_∞-stability of the error dynamics (3.19).

Definition 3.1 *The dynamical system $e_{k+1} = f(k, e_k, q_k)$ is said globally uniformly l_∞-stable with performance level γ if the following conditions are satisfied [75]:*

(1) The disturbance-free system (i.e., $q_k = 0$) for all $k \geq 0$) is globally uniformly exponentially stable with respect to the origin.

(2) For a zero initial condition(i.e., $e_{k_0} = 0$), and every bounded unknown input q_k, $||e_k|| \leq \gamma ||q_k||_\infty$ for all $k \geq 0$.

(3) For every initial condition $e_{k_0} = e_0$, and every bounded unknown input q_k, $\lim_{k\to\infty} \sup ||e_k|| \leq \gamma ||q_k||_\infty$.

Lemma 3.1 *Consider the dynamical system $e_{k+1} = f(k, e_k, q_k)$. If there exists a function $V : \mathbb{R}^n \mapsto \mathbb{R}$, and scalars $\alpha \in (0, 1)$, $\beta_1, \beta_2 \geq 0$ and $\mu_1, \mu_2 \geq 0$ such that $\forall\, k \geq 0$:*

$$\beta_1 ||e_k||^2 \leq V(e_k) \leq \beta_2 ||e_k||^2 \tag{3.20}$$

and

$$V(e_{k+1}) - V(e_k) \leq -\alpha(V(e_k) - \mu_1 ||q_k||^2) \tag{3.21}$$

$$||e_k||^2 \leq \mu_2 V(e_k) \tag{3.22}$$

then the system is globally uniformly l_∞-stable with performance level $\gamma = \sqrt{\mu_1 \mu_2}$ with respect to output disturbance q_k [75]. □

Theorem 3.1 *Consider the error dynamics (3.19) and $\alpha \in (0, 1)$. If there exist $P \in \mathbb{R}^{n\times n}$, $P = P^T \geq 0$ and $Z \in \mathbb{R}^{n\times p}$ such that the following LMI admits a solution*

$$\begin{bmatrix} -P & * \\ \Omega_{21} & \Omega_{22} \end{bmatrix} \leq 0, \tag{3.23}$$

where

$$\Omega_{21} = \begin{bmatrix} A^T P - C^T Z^T \\ E^T P - D^T Z^T \end{bmatrix}, \Omega_{22} = \begin{bmatrix} -(1-\alpha)P & 0_{n\times(n+p)} \\ 0_{(n+p)\times n} & -\alpha I \end{bmatrix}$$

then the error dynamics (3.19) with gain $L = P^{-1}Z$ are l_∞-stable with performance level $\gamma = 1/\sqrt{\lambda_{min}(P)}$.

Proof 3.1 *Let $\mathcal{H} = A - LC$ and $\mathcal{F} = E - LD$. By considering a quadratic candidate Lyapunov function $V(e_k) = e_k^T P e_k$, and let $\Delta_k = V(e_{k+1}) - V(e_k)$, we have that:*

$$\begin{aligned} \Delta_k = e_k^T(\mathcal{H}^T P \mathcal{H} - P)e_k + q_k^T(\mathcal{F}^T P \mathcal{F})q_k \\ + e_k^T(\mathcal{H}^T P \mathcal{F})q_k + q_k^T(\mathcal{F}^T P \mathcal{H})e_k \end{aligned} \tag{3.24}$$

Then, by considering $\mu_1 = 1$, the condition (3.21) *defines a matrix inequality*

$$\Delta_k + \alpha(V(e_k) - ||q_k||^2) \leq 0 \tag{3.25}$$

which, by resorting to simple manipulations, can be rewritten as the following bilinear matrix inequality

$$\begin{bmatrix} \mathcal{H}^T P \mathcal{H} - (1-\alpha)P & \mathcal{H}^T P \mathcal{F} \\ \mathcal{F}^T P \mathcal{H} & \mathcal{F}^T P \mathcal{F} - \alpha I \end{bmatrix} \leq 0 \tag{3.26}$$

Now, let $Z = PL$, the above can be re-arranged as

$$\Omega_{21} + \Omega_{22} P^{-1} \Omega_{22}^T \leq 0 \tag{3.27}$$

where

$$\Omega_{21} = \begin{bmatrix} A^T P - C^T Z^T \\ E^T P - D^T Z^T \end{bmatrix} \quad \Omega_{22} = \begin{bmatrix} -(1-\alpha)P & 0_{n\times(n+p)} \\ 0_{(n+p)\times n} & -\alpha I \end{bmatrix}$$

By applying a Schur complement, (3.27) *is equivalent to the LMI* (3.23)*. Moreover, as shown in [75], the conditions* (3.20) *and* (3.22) *are satisfied if $\beta_1 = \lambda_{min}(P), \beta_2 = \lambda_{max}(P)$ and $\mu_2 = 1/\lambda_{min}(P)$. As a consequence, according to Lemma 3.1, the state error dynamics (3.19) are l_∞-stable with performance level $\gamma = 1/\sqrt{\lambda_{min}(P)}$, concluding the proof.* □

3.4.2 Input Disturbance Estimation

The estimation of the output disturbance vector q_k can be obtained by multiplying both sides of the output equation of (3.15) by D^+ and replacing x_k with $\hat{x}_k$, i.e.,

$$\hat{q}_k = D^+ \hat{y}_k - D^+ C \hat{x}_k \tag{3.28}$$

To prove l_∞-stability of the unknown term $\hat{q}_k$, we can consider the estimation error $e_k^q = q_k - \hat{q}_k$. Then, we can write that

$$e_k^q = -D^+ C e_k \tag{3.29}$$

Since according to theorem 3.1, $\lim_{k\to\infty} \sup ||e_k|| \leq \gamma ||q_k||_\infty$, then the steady-state error e_k^q enjoys the following bound

$$\begin{aligned} \lim_{k\to\infty} \sup ||e_k^q|| &\leq ||D^+C||\gamma||q_k||_\infty \\ &\leq ||D^+C||\sqrt{\mu_2}||q_k||_\infty \end{aligned} \tag{3.30}$$

Remark 3.3 *If w_k and η_k do not follow a bounded distribution, then an extension of the results in section 3.4.1 and 3.4.2 is needed to ensure globally uniformly l_∞-stability of the error dynamics* (3.19) *and bounded error in the estimation of the unknown vector q_k.*

3.5 Anomaly Detector Design

In this section, first, we design a Gaussian anomaly detector leveraging the above-developed encoding scheme. Then, to prevent the existence of intelligent undetectable attacks (e.g., replay and covert attacks), we enhance the detection capabilities by adding randomness to the used encoding mechanism.

3.5.1 Gaussian Anomaly Detector

The estimation of the cumulative output disturbance vector v_k can be obtained from $\hat{q}_k$ as $\hat{v}_k = D\hat{q}_k$ (see (3.14), (3.28) and (3.30)). Therefore, $\hat{v}_k$ is a stochastic variable with a bounded distribution. However, since the lower and upper bounds on each component of $\hat{v}_k$ depend on the signal y_k, to design the anomaly detector, we here assume that $\hat{v}_k$ can be described by a general unbounded Gaussian distribution, whose Probability Density Function (PDF) is

$$f(\hat{v}_k, \mu, \Sigma) = \frac{1}{(|\Sigma|^{\frac{1}{2}})(2\pi)^{\frac{p}{2}}} e^{-\frac{1}{2}(\hat{v}_k - \mu)^T \Sigma^{-1} (\hat{v}_k - \mu)} \tag{3.31}$$

where $\mu \in \mathbb{R}^p$, $\Sigma \in \mathbb{R}^{p\times p}$ are the mean and covariance of the distribution, respectively.

Remark 3.4 *If the lower and upper bound for each component of $\hat{v}_k$ are estimated from the available data samples (e.g., by using the Extreme Studentized Deviation method [76]),*

then it is possible to use the maximum likelihood method to estimate a truncated distribution (e.g., a truncated Gaussian distribution [77]) instead of (3.31). □

By collecting, in an attack-free scenario, N samples of $\hat{v}_k$, an estimation of μ and Σ can be obtained by maximizing the likelihood function [78]. In particular, for the considered distribution, we have that:

$$\begin{aligned} \mu &= \frac{1}{N}\sum_{i=1}^{N}\hat{v}_k^{(i)} \\ \Sigma &= \frac{1}{N}\sum_{i=1}^{N}(\hat{v}_k^{(i)}-\mu)(\hat{v}_k^{(i)}-\mu)^T \end{aligned} \tag{3.32}$$

Therefore, given (3.31), a binary anomaly detector can be designed to trigger an anomaly if the likelihood associated to new samples $\hat{v}_k$ is lower of a given threshold $\zeta > 0$ (see Algorithm 3.1). The tuning of ζ can be obtained from the desired false alarm rate $\varepsilon > 0$. Define the false alarm probability of the binary anomaly detection rule $f(\hat{v}_k, \mu, \Sigma) < \zeta$ on the available N samples as

$$f_{alarm} = \frac{FP}{FP + TN} \tag{3.33}$$

where FP is the number of times an anomaly is detected in the absence of attacks (i.e., $f(\hat{v}_k, \mu, \Sigma) < \zeta$) and TN is the number of times an anomaly is not detected in the absence of attacks (i.e., $f(\hat{v}_k, \mu, \Sigma) \geq \zeta$). Then, the hyper-parameter ζ can be found using, e.g., a *grid or random* search method [79].

Algorithm 3.1 Anomaly Detector

Input: $\hat{v}$, ζ
Output: anomaly $= \{0, 1\}$

1: Compute the likelihood of $\hat{v}$'s new samples:
$$f(\hat{v}_k, \mu, \Sigma) = \frac{1}{(|\Sigma|^{\frac{1}{2}})(2\pi)^{\frac{p}{2}}} e^{-\frac{1}{2}(\hat{v}_k-\mu)^T\Sigma^{-1}(\hat{v}_k-\mu)}$$
2: **if** $f(\hat{v}_k, \mu, \Sigma) \leq \zeta$ **then** anomaly $= 1$
3: **else** anomaly $= 0$
4: **end if**

Remark 3.5 *Please note that the used encoding mechanism prescribes the use of a static transformation matrix Φ_d in* (3.10). *Therefore, if the attacker is aware of the plant model*

(i.e., A, B, C) and Φ_d, the proposed detector won't be able to detect covert attacks [13, 21]. Moreover, steady-state replay attacks will also be undetectable [20]. □

Motivated by the above, in the next section, we properly change the encoding mechanism to allow the detector to detect intelligent coordinated attacks.

3.5.2 Time-Varying Dimensionality Reduction Scheme

This section introduces a time-varying (random) encoding scheme to prevent the attacker from performing coordinated stealthy attacks (see Remark 3.5).

Assumption 3.1 *Two identical and cryptographically secure pseudo-random number generators [80] are available on both the encoder and decoder. Moreover, a secret common seed is offline shared between them.*

The proposed solution, which recalls the idea in [81], prescribes that the encoder performs three main operations:

- First, the encoding matrix $\Phi_d, d \leq p$ is chosen randomly from $\mathcal{T}$ in (3.9). By denoting with Φ_{d_k} the value of Φ_d at the time k, we have $\tilde{y}_k = \Phi_{d_k}^T y_k$,
- Second, the vector $\tilde{y}_k \in \mathbb{R}^d$ is extended with an auxiliary random vector, namely $y_k^e := [y_k^{1,e}, y_k^{2,e}, \ldots, y_k^{p-d,e}] \in \mathbb{R}^{p-d}$. Therefore, the augmented vector $y_k^{aug} := [\tilde{y}_k^T, y_k^{eT}]^T$ is constructed.
- Third, a random permutation matrix, namely $\Omega_k \in \mathbb{R}^p$, is applied on the augmented output vector y_k^{aug}, to obtain $\tilde{y}_{r_k}$, i.e.,

$$\tilde{y}_{r_k} = \Omega_k \underbrace{\begin{bmatrix} \tilde{y}_k \\ y_k^e \end{bmatrix}}_{y_k^{aug}} \tag{3.34}$$

  which is transmitted over the network.

On the other hand, the decoder, performs the inverse operations to recover $\hat{y}_k$:

- First, the random matrices Φ_{d_k} and Ω_k are randomly generated.

- Second, the received augmented signal, namely y_k^{aug} is determined, i.e.,

$$y_k^{aug} = {\Omega_k}^{-1} \tilde{y}_{r_k} \tag{3.35}$$

- Third, the first d components of y_k^{aug}, namely $\tilde{y}_k$, are multiplied by Φ_{d_k}, i.e.,

$$\hat{y}_k = \Phi_{d_k} \tilde{y}_k \tag{3.36}$$

Remark 3.6 *Please note that given Assumption 3.1, the random matrices Φ_{d_k} and Ω_k generated at the encoder's and decoder's sides are identical.* □

Remark 3.7 *It is important to remark that the above-presented randomization step for the sensor measurements cannot be used alone (i.e., with $d = p$ and $\Phi_p = I_p$) to prevent the existence of stealthy attacks. Indeed, if the sensor measurements have different natures (e.g., different range values), then an attacker with knowledge about the physical system dynamics might be able to discriminate the sensor components regardless of their position in the measurement vector $\tilde{y}_{r_k}$. On the other hand, the proposed dimensionality reduction can be, in principle, used alone (i.e., with $\Phi_{d_k} \in \mathcal{T}$ randomly changing, $y_k^e = \text{‰}$ and $\Omega_k = I_d$). However, in this case, the attacker, just observing the dimension of the transmitted sensor vector $\tilde{y}_k$ might be aware of how many principle components have been used. Therefore, it can leverage this information along with the fact that the set of admissible encoding matrices Φ_{d_k} is fixed and constant (i.e., the set $\mathcal{T}$ does not change over time) to identify the used encoding mechanism. Therefore, the combined use of an encoding mechanism and randomization is desirable to make the proposed randomization mechanism more secure.* □

3.5.3 Absence of Stealthy Attacks

In what follows, we prove that given the proposed detection scheme, the estimated output disturbance vector $\hat{q}_k$ in the absence of attacks is always different from the one obtained in the presence of any FDI attacks, so preventing the existence of well-known stealthy attacks w.r.t. the proposed anomaly detector (see Algorithm 3.1). In particular,

the effectiveness is proved for three well-studied attacks, namely (1) FDI on the sensor measurements, (2) Replay attack [20], and (3) Covert attack [21].

FDI Attacks On The Sensor Measurements

By encoding the sensor measurement vector according to (3.34), the transmitted signal will be $\tilde{y}_{r_k}$. If the attacker injects y_k^a, then the decoder receives

$$y_k^{'} = \tilde{y}_{r_k} + y_k^a \tag{3.37}$$

Based on (3.35)-(3.36), we have that

$$y_k^{'aug} = {\Omega_k}^{-1} y_k^{'} = {\Omega_k}^{-1}(\tilde{y}_{r_k} + y_k^a) = y_k^{aug} + {\Omega_k}^{-1} y_k^a \tag{3.38}$$

$$y_k^{'} = \Phi_{d_k}\tilde{y}_k + \Phi_{d_k}[{\Omega_k}^{-1} y_k^a]_{[1:d]} \tag{3.39}$$

where, $[\cdot]_{[1:d]}$ is an operator that selects the first d components of its argument. Let $\hat{q}_k^{'}$ be the estimation of q_k in the presence of attack, then, by means of simple manipulations

$$\hat{q}_k^{'} = D^{+}(y_k^{'} - C\hat{x}_k) \tag{3.40}$$

and

$$\hat{q}_k^{'} = D^{+}(\hat{y}_k + \Phi_{d_k}[{\Omega_k}^{-1} y_k^a]_{[1:d]} - C\hat{x}_k) \tag{3.41}$$

Note that in the above Ω_k and Φ_{d_k} randomly change and that the null space of Φ_{d_k} is unknown to the attacker. Consequently, sensor measurement attacks are undetectable (irrespective of ζ, Ω_k and Φ_{d_k}) iff $\hat{q}_k^{'} - \hat{q}_k = 0, \forall\, k$, which implies that the only guaranteed stealthy attack vector is $y_k^a = 0, \forall\, k$. □

Covert Attacks

In the presence of a covert attack, the decoded signal received by the control center is:

$$y_k^{'} = \hat{y}_k + y_k^{u^a} \tag{3.42}$$

where

$$y_k^{u^a} = \Phi_{d_k}(\Phi_{d_k}^T C \sum_{j=0}^{k-1}(A^j B u_{k-1-j}^a))$$

is related to the input attack u_k^a. Moreover, as shown by (3.40), the estimation of q_k will be

$$\hat{q}_k^{'} = D^+(\hat{y}_k + y_k^{u^a} - C\hat{x}_k) \tag{3.43}$$

However, the effect of u_k^a on the encoded sensor measurements $\tilde{y}_{r_k}$ is

$$y_{r_k}^{u^a} = \Omega_k \left[\left(\Phi_{d_k}^T C \sum_{j=0}^{k-1}(A^j B u_{k-1-j}^a) \right), 0_{p-d}^T \right]^T \tag{3.44}$$

Therefore, covert attacks are stealthy irrespective of detector's threshold ζ iff the attacker is capable of exactly computing (3.44) $\forall\, k$ and subtract it from $\tilde{y}_{r_k}$ by means of the FDI attack $y_k^a = -y_{r_k}^{u^a}$. However, in the proposed architecture, both Ω_k and Φ_{d_k} are unknown to the attacker. As a consequence, (3.44) cannot be computed, and the attack is not stealthy. □

Replay Attacks

If the attacker performs the replay attack $\tilde{y}_k = \tilde{y}_{k-\tau}, \tau > 1$, the vector $\hat{y}_k$ received by the control center become

$$\hat{y}_k = \Phi_{d_k}\Omega_k^{-1}\Omega_{k-\tau}\Phi_{d_{k-\tau}}^T y_{k-\tau}$$

and the estimation of the output disturbance in the presence of a replay attack can be written as

$$\hat{q}_k^{'} = D^+(\Phi_{d_k}\Omega_k^{-1}\Omega_{k-\tau}\Phi_{d_{k-\tau}}^T y_{k-\tau} - C\hat{x}_k) \tag{3.45}$$

Consequently, since Ω_k and Φ_{d_k} randomly change at each k, replay attacks cannot guarantee that $\hat{q}_k^{'} - \hat{q}_k = 0, \forall\, k$. □

Remark 3.8 *The above analysis does not exclude the possibility that other ad-hoc undetectable FDI attacks exist against the proposed detection mechanism. For example, an attack might be able to inject a sufficiently small FDI injection such that the likelihood of $\hat{v}_k$ is still*

within the used threshold. The characterization of this class of attacks and its impact on the plant's performance can be obtained by exploiting an analysis similar to the one proposed in, e.g., [82], and it is left for future studies. □

3.6 Simulation Results

In this section, simulation results are shown to testify to the effectiveness of the proposed strategy. To this end, the quadruple-tank water system described in [83] has been used as a testbed. The system consists of four tanks, where the water levels $h_i, i = 1, \ldots, 4$, are the state components of the system, i.e., $x = [h_1, \ldots, h_4]$ while the two valves v_1 and v_2 are the control inputs, i.e., $u = [v_1, v_2]^T$. Four sensors are available to measure the water levels in all the tanks (see Figure 3.3). The plant dynamics have been discretized with a sampling

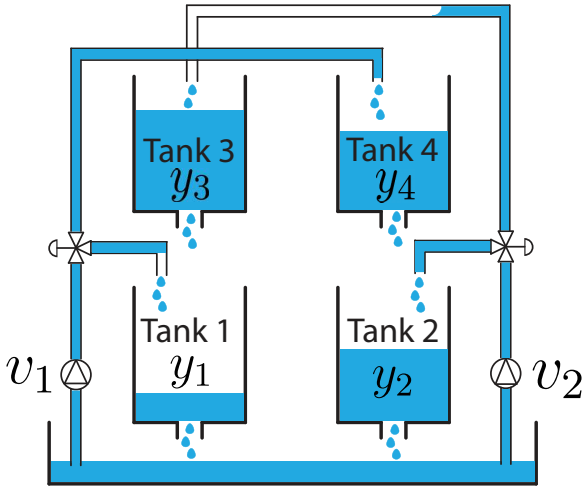


Figure 3.3: Quadruple Water Tank System.

time $T_s = 1\,\text{sec}$ and linearized around the operating equilibrium point

$$x_{eq} = \begin{bmatrix} 5 & 5 & 2.044 & 1.399 \end{bmatrix}^T, u_{eq} = \begin{bmatrix} 0.724 & 1.165 \end{bmatrix}^T.$$

The linearized system matrices are

$$A = \begin{bmatrix} 0.975 & 0 & 0.042 & 0 \\ 0 & 0.977 & 0 & 0.044 \\ 0 & 0 & 0.958 & 0 \\ 0 & 0 & 0 & 0.956 \end{bmatrix}, B = \begin{bmatrix} 0.0515 & 0.0016 \\ 0.0019 & 0.0447 \\ 0 & 0.0737 \\ 0.0850 & 0 \end{bmatrix},$$

$$C = \begin{bmatrix} 0.2 & 0 & 0 & 0 \\ 0 & 1.5 & 0 & 0 \\ 0 & 0 & 10.5 & 0 \\ 0 & 0 & 0 & 1.7 \end{bmatrix}, E = \begin{bmatrix} I_4 & 0_4 \end{bmatrix}, D = \begin{bmatrix} 0_4 & I_4 \end{bmatrix}$$

Moreover, the considered process and measurement noises are $w_k \sim \mathcal{N}(0, 10^{-4})$, and $\eta_k \sim \mathcal{N}(0, 10^{-4})$.

The networked control system is equipped with a state feedback controller in charge of stabilizing the system around the equilibrium point. The controller gain is:

$$K = \begin{bmatrix} -2.0303 & -0.0657 & -0.0412 & -3.2652 \\ -0.0716 & -2.3495 & -3.7746 & -0.0535 \end{bmatrix}$$

The UIO is designed with $\alpha = 0.95$, and the resulting observer gain L is

$$L = \begin{bmatrix} 4.8750 & 0 & 0.0040 & 0 \\ 0 & 0.6513 & 0 & 0.0259 \\ 0 & 0 & 0.0912 & 0 \\ 0 & 0 & 0 & 0.5624 \end{bmatrix}$$

The auxiliary random vectors y_k^e have been generated follows a normal Gaussian distribution. To show the effectiveness of the proposed UIO, Fig. 3.4 shows the results of a simulation experiment where an output disturbance $v_k = 1.5\sin(k)$ for $250 \leq k \leq 350$ has been injected into the sensor measurements (in the absence of attacks). Specifically, the proposed UIO can estimate the unknown output disturbance v_k with an average error norm-2 error, $\|v_k - \hat{v}_k\|_2$, equals to 0.148. To design the proposed time-varying encoding scheme,

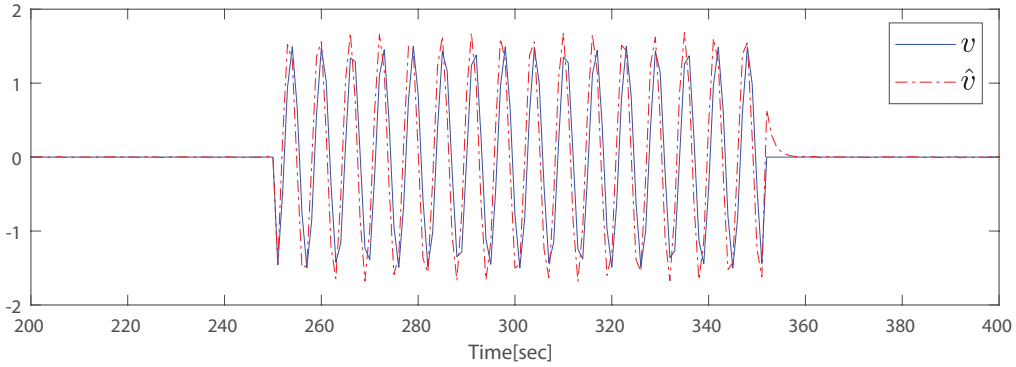


Figure 3.4: Output disturbance and its estimation using the proposed UIO.

we have collected 1000 samples of the sensor measurements in the absence of attacks. Then, by solving the optimization (3.6), the following encoding matrix ϕ is obtained:

$$\phi = \begin{bmatrix} -0.09 & 0.69 & -0.71 & 0.08 \\ -0.62 & 0.32 & 0.32 & -0.62 \\ 0.32 & 0.64 & 0.61 & 0.32 \\ 0.70 & 0.09 & -0.08 & -0.70 \end{bmatrix}$$

Figs. 3.5 shows the *modus operandi* of the proposed control architecture (i.e., number of PCs used at each time instant, state-estimation error, disturbance estimation error) in the absence of attacks, starting from $x(0) = \begin{bmatrix} 5 & 5 & 2 & 1 \end{bmatrix}^T$ and for $100 \leq k \leq 400$. Moreover, Fig 3.6 shows the time evolution of y_k and $\tilde{y}_k$. It is possible to note that although the system is at the equilibrium point (i.e., the vector y_k is constant over time), the transmitted encoded measurements (i.e., $\tilde{y}_k$) are not. The latter is justified by the fact that the encoding matrix Φ_{d_k} (see Fig. 3.5) and Ω_k are changed at every time step. The latter is a key feature of the proposed architecture that, as shown in the next subsections, allows the detection of different FDIs. Moreover, we have estimated the Gaussian distribution parameters (3.31) as in (3.32). Moreover, by using as desired false alarm rate $\varepsilon = 3\%$, the resulting anomaly threshold, normalized with respect to the maximum value of (3.31), is $\zeta = 0.14$. In what follows, to evaluate the performance of the proposed anomaly detector against the

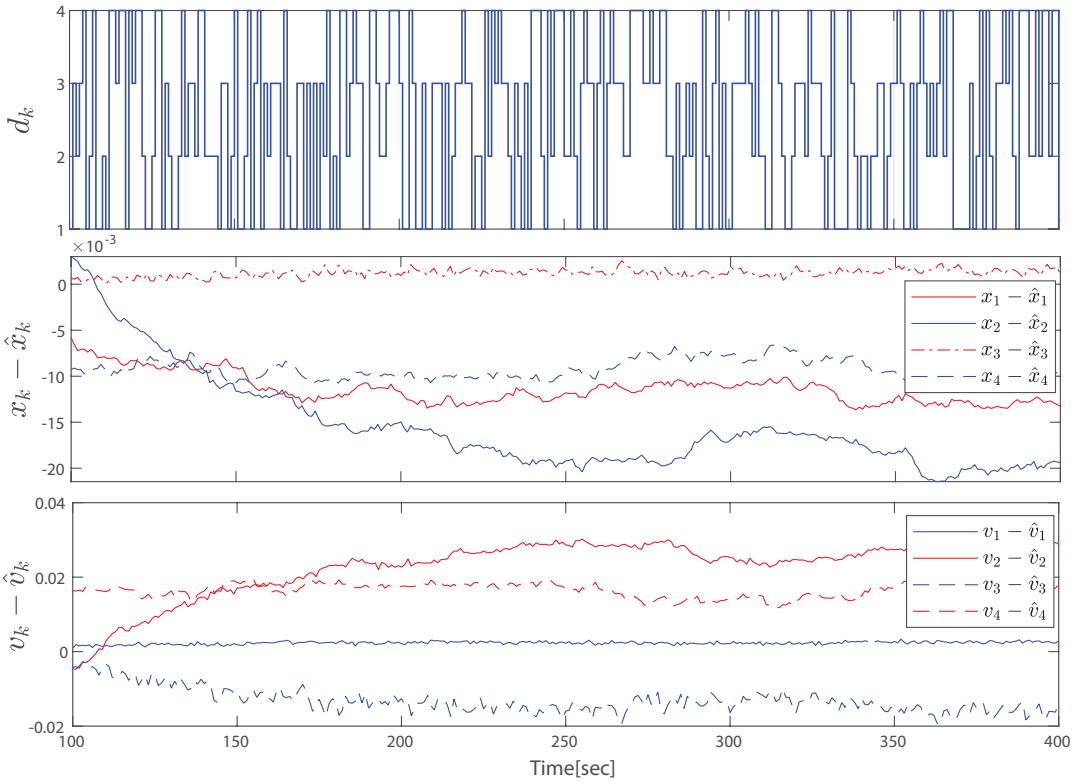


Figure 3.5: Modus operandi of the proposed control architecture in the absence of attacks (number of PCs, state-estimation error, disturbance estimation error).

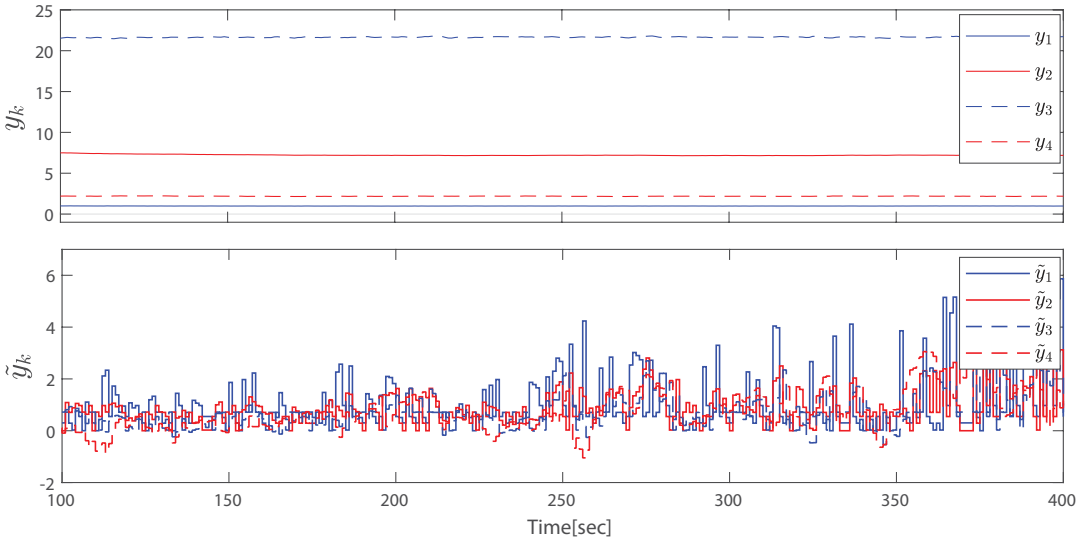


Figure 3.6: Top plot shows the original sensor measurements, y_k. Bottom plot shows the encoded sensor measurements, $\tilde{y}_k$.

considered attacks, the following F_{score} formula is used [84]:

$$F_{score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{3.46}$$

where $0 \leq F_{score} \leq 1$, TP is the number of times an anomaly is detected during the attack and FN is the number of times an anomaly is not detected during the attack. In what follows, the detection performance of the proposed architecture is evaluated in the presence of replay and covert attacks and contrasted with the solution in [1].

3.6.1 Replay Attack

We consider a replay attack affecting the system starting from a steady-state condition. In particular, the attacker records the encoded sensor measurements $\tilde{y}_k$ for $150 \leq k < 250$, and replay the recorded information to the decoder for $250 \leq k < 350$, i.e. $\tilde{y}_k = y_k^a = \tilde{y}_{k-100}$. During the same time frame, the attacker injects the following signal on the actuation channel:

$$u_k^a = \begin{cases} \begin{bmatrix} 0 & 0 \end{bmatrix}^T, & \text{if } k < 250 \\ \begin{bmatrix} k/50 & k/150 \end{bmatrix}^T, & \text{if } 250 \leq k \leq 350 \\ \begin{bmatrix} 0 & 0 \end{bmatrix}^T & \text{if } k \geq 350 \end{cases} \tag{3.47}$$

Fig 3.7 shows that the proposed encoding scheme allows the Gaussian-based detector to discover the presence of replay attacks. Indeed, although the system is in a steady-state condition when the attack starts, the replied sensor measurement signal won't be correctly decoded. This is justified by the fact that the randomly changing encoding and permutation matrices used at the time k are different from the ones embedded into the replied vector $\tilde{y}_{k-100}$ at the time $k - 100$. As a consequence, the likelihood of $\hat{v}_k$ becomes lower than the threshold. In this experiment, for $k \in [100, 350]$, the detector F_{score} is 0.95.

3.6.2 Covert Attack

We consider a covert-attack where the input attack vector is as in (3.47) and the output attack vector y_k^a is computed as prescribed by (2.8). Fig. 3.8 shows the likelihood normalized with respect to the maximum value of $f(\hat{v}_k, \mu, \Sigma)$. It is possible to appreciate that as soon as the attack is started, the likelihood signal goes below the used threshold, allowing the detection of the attack. Similar to what was discussed for the replay attack, this is justified by the fact that the randomly changing permutation and encoding matrices do not allow

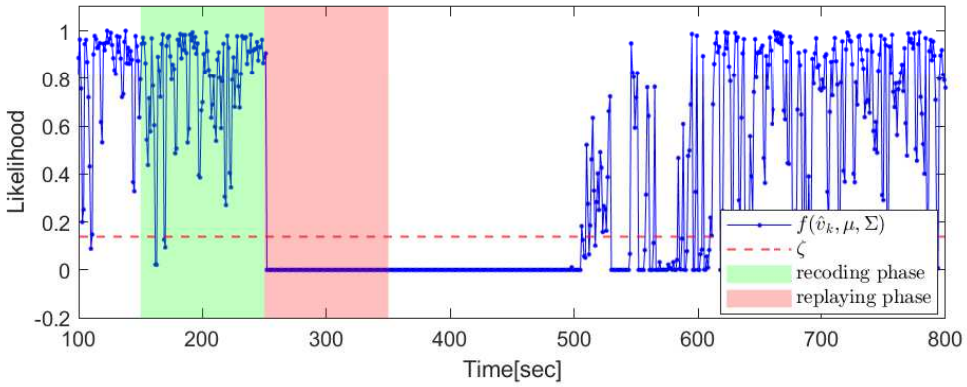


Figure 3.7: Water-tank under replay attack. Normalized likelihood of $\hat{v}_k$.

the attacker to compute the stealthy attack (2.8). In this experiment, for $k \in [100, 350]$, the detector F_{score} is 0.92.

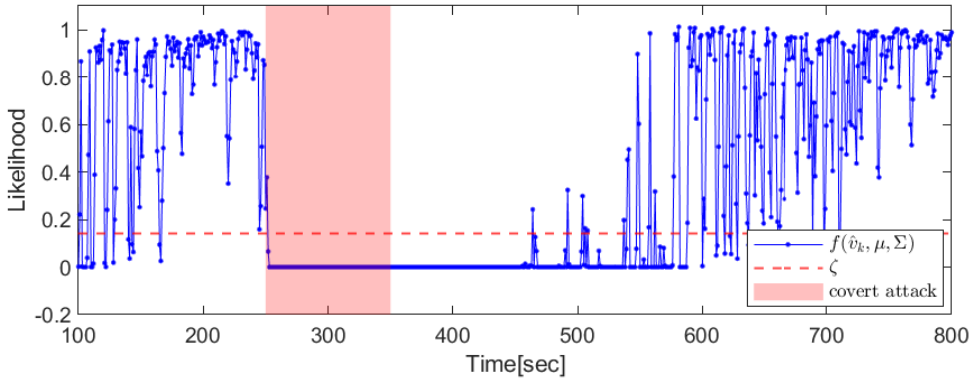


Figure 3.8: Water-tank under covert attack. Normalized likelihood of $\hat{v}_k$.

3.6.3 Comparison

Here, we compare the proposed solution with the one in [1]. First, by considering the quadruple water tank system and the replay and covert attacks discussed in the previous subsections, we contrasted the detection performance of the two encoding schemes. For such attacks affecting both communication channels, we configured the competitor encoding scheme to randomly select the encoding scheme in a set of 10 invertible encoding matrices (different from the identity). Such a set has been selected to obtain the best possible detection result for [1]. The detection F_{score} results for both attacks is reported in Table 3.1. It shows that the proposed scheme presents slightly better detection performance. However, it is essential to recall that the proposed solution has another advantage. In particular, the

combined use of an encoding mechanism with a randomly changing permutation matrix does not allow the attacker to discriminate among the transmitted measurements, preventing the attacker from estimating the encoding matrices. Moreover, we have compared the performance of the two detection schemes for the unstable 2D-LTI system and sensor measurement attack considered in [1, Section VI.A]. To comply with the results shown in [1], we define the residual signal $y_k - C\hat{x_k}$ without and with the attack as z_k and $z_k^{'}$, respectively. Moreover, we denote with Δz_k and $\Delta \tilde{z}_k$ the difference between such residual signal without and with the encoding mechanism. Fig. 3.9 shows that in the presence of the sensor attack and in the absence of encoding, the signal $||\Delta z_k||_2$ remains bounded. On the other hand, in the presence of the encoding mechanisms, the same signal increases over time. Consequently, we can conclude that the proposed encoding scheme can detect the stealth attacks considered in [1]. The effectiveness of the proposed solution is confirmed by the results in Fig. 3.10, where it is shown that the likelihood of $\hat{v}_k$ is below the detection threshold $\zeta = 0.2$ (obtained for a 3% false alarm rate) when the proposed encoding mechanism is active.

Table 3.1: F_{score}: proposed solution vs encoding based on [1].

Attack	proposed solution	encoding based on [1]
Replay attack	0.95	0.88
Covert attack	0.92	0.85

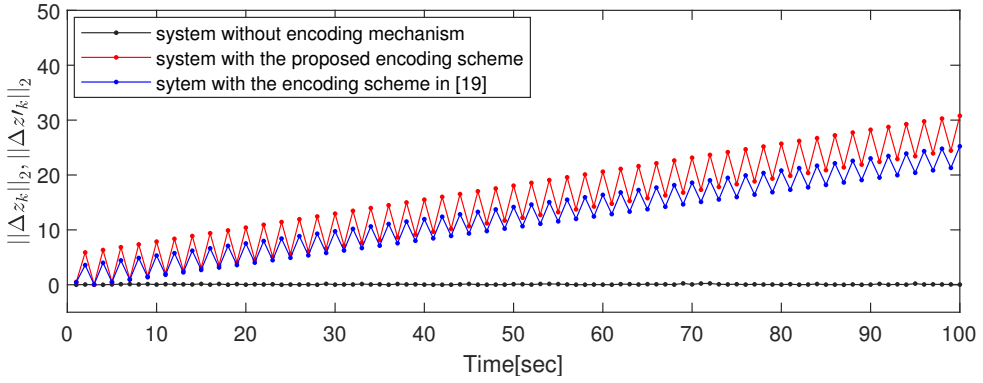


Figure 3.9: Unstable 2D-LTI system under the FDI attack described in [1]. $||\Delta z_k||_2$ and $||\Delta \tilde{z}_k||_2$ signals.

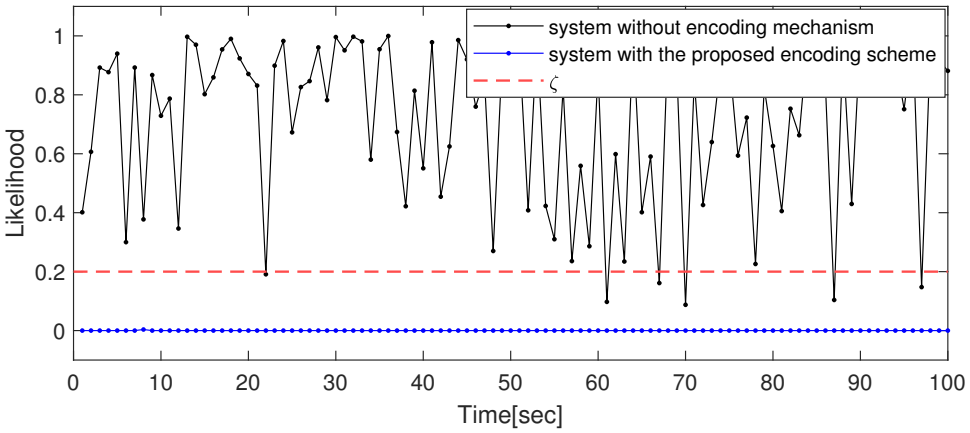


Figure 3.10: Unstable 2D-LTI system under the FDI attack described in [1]. Normalized likelihood of $\hat{v}_k$.

3.7 Conclusion and Future Works

In this chapter, a networked control architecture capable of preventing the existence of stealthy cyber-attacks has been presented. The proposed solution leverages the principal components of the sensor measurements to send to the control center a modified randomly changing signal containing a subset of the principal components of the measurements. It has been shown that such an architecture has two main benefits. On one side, it drastically limits the attacker's disclosure and disruptive resources, nullifying the attacker's possibility of producing stealthy false-data injections. On the other hand, it allows designing a decoder on the control center to reconstruct the original measurement vector optimally. Moreover, by considering the resulting bounded decoding error, an ad-hoc tuned UIO has been designed to simultaneously estimate the system's states and all the unknown signals. The estimated unknown inputs have been leveraged to design a Gaussian anomaly detector. Simulation and comparison results have shown the effectiveness of the proposed architecture against different intelligent attacks. Future studies will be devoted to enhancing the proposed architecture with attack mitigation strategies that are activated once an attack is revealed.

Chapter 4

Data-Driven Robust Backward Reachable Sets for Set-Theoretic Model Predictive Control

4.1 Introduction

In this chapter, we propose a novel approach for computing robust backward reachable sets from noisy data for unknown constrained linear systems subject to bounded disturbances. We develop an algorithm for obtaining zonotopic inner approximations that can be used for control purposes. It is shown that such sets, if built on an extended space including states and inputs, can be used to embed the system's one-step evolution in the computed extended region. Such a result is then exploited to build a set-theoretic model predictive controller that, offline, builds a recursive family of robust data-driven reachable sets and, online, computes recursively admissible control actions without explicitly resorting to either a model of the system or the available data. The validity of the proposed data-driven solution is verified by means of a numerical simulation and its performance is contrasted with the model-based counterpart.

4.1.1 Contribution of the Work

To the best of the author's knowledge, the problem of computing robust Backward Reachable Sets (BRS)[1] from a collection of noisy input-state trajectories has not yet been explored in the literature. In this regard, this chapter extends the data-driven approach developed in [48] (computing over approximations of forward reachable sets) to compute data-driven zonotopic inner approximations of robust backward reachable sets for unknown linear systems subject to bounded state and input constraints as well as disturbances. We develop a novel augmented description of BRS, which is then used to obtain a novel data-driven implementation of the Set-Theoretic MPC (ST-MPC) controller developed in [57]. The proposed solution borrows from [48] only the characterization, via a matrix zonotope, of all the linear models consistent with the available data. Starting from this common point, the here proposed methodology to compute BRS and a data-driven ST-MPC are different and provide solutions to challenges not faced in [48]:

- Zonotopes not closed under Minkowski difference and set intersection operations required to compute BRS, and
- Exponential number of constraints in the data-driven implementation of ST-MPC. In addition, while the data-driven MPC proposed in [85] is based on the online computation of FRS for the required prediction horizon, the proposed data-driven ST-MPC resorts to a family of offline computed BRS. Consequently, the proposed controller is capable of moving offline most of the required computations, leaving online an MPC problem with a prediction horizon equal to one.

A Matlab implementation of the here-developed algorithm is available on the following GitHub page: https://github.com/PreCyseGroup/Data-Driven-ST-MPC.

4.2 Preliminaries and Problem Formulation

Lemma 4.1 *[86, Theorem 2] Consider $\mathcal{Z}(c, G)$, with p independent generators, $c \in \mathbb{R}^n, G \in \mathbb{R}^{n\times p}$. Let $v = \binom{p}{n-1}$ be the number of combinations of $n-1$ distinct generators, and*

[1] Also known as robust one-step controllable sets

$\mathcal{I}_i = \{\delta_1^i, \ldots, \delta_{n-1}^i\}$ *the column indices associated to each* $i-th$ *combination. The polytopic* $\mathcal{H}-$*representation of* $\mathcal{Z}(c,G)$ *is* $\mathcal{Z}(c,G) = \{x \in \mathbb{R}^n : Cx \leq d\}$, *where*

$$\begin{aligned} C &= \begin{bmatrix} \bar{C} \\ -\bar{C} \end{bmatrix}, \bar{C} = \begin{bmatrix} \bar{C}_1 \\ \vdots \\ \bar{C}_v \end{bmatrix}, \bar{C}_i = \frac{CP_n(G^{<\mathcal{I}_i>})^T}{\|CP_n(G^{<\mathcal{I}_i>})\|_2}, \\ d &= \begin{bmatrix} \bar{C}c + \Delta d \\ -\bar{C}c + \Delta d \end{bmatrix}, \Delta d = \sum_{v=1}^{p} |\bar{C}g^{(v)}|. \end{aligned} \tag{4.1}$$

and $G^{<\mathcal{I}_i>}$ *contains the column in* G *specified by* $\mathcal{I}_i$.

Remark 4.1 *A zonotope (matrix zonotope) can always be represented as the convex hull of its vertices (matrix vertices) [61, Sec. 3.3.3], [87].*

4.2.1 Constrained Plant Model

Consider the class of discrete-time linear time-invariant systems subject to a bounded but unknown process disturbance and described by

$$x_{k+1} = Ax_k + Bu_k + w_k \tag{4.2}$$

where $k \in \mathbb{Z}_+ = \{0, 1, ...\}$ and $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}$ are the system matrices. Moreover, $x_k \in \mathbb{R}^n$, $u_k \in \mathbb{R}^m$, and $w_k \in \mathbb{R}^n$ are the state, control, and disturbance vectors, respectively. In addition, it is prescribed that:

$$x_k \in \mathcal{X} \subset \mathbb{R}^n, \quad u_k \in \mathcal{U} \subset \mathbb{R}^m, \quad w_k \in \mathcal{W} \subset \mathbb{R}^n \tag{4.3}$$

where $\mathcal{X}, \mathcal{U}, \mathcal{W}$ are convex, compact, and contain the origin.

Definition 4.1 *Consider* (4.2)-(4.3) *and a target set* $\mathcal{T}^{j-1} \subseteq \mathcal{X}$. *The set of states Robust One-Step Controllable (ROSC) to* $\mathcal{T}^{j-1}$ *is*

$$\mathcal{T}^j = \{x \in \mathcal{X} : \exists u \in \mathcal{U} : x^+ \in \mathcal{T}^{j-1}, \forall w \in \mathcal{W}\} \tag{4.4}$$

and $x^+ := Ax + Bu + w$.

4.2.2 Problem Statement

Assumption 4.1 *The matrices* A, B *in* (4.2) *are unknown, and the disturbance set* $\mathcal{W}$ *can be represented (or over-approximated) as a zonotope described in the* $\mathcal{G}-$ *or* $\mathcal{H}-$ *representation, i.e.,*

$$\mathcal{W} \quad = \quad \mathcal{Z}_w(c_w, G_w) = \{w \in \mathbb{R}^n : H_w w \leq h_w\}, \tag{4.5}$$

where $c_w \in \mathbb{R}^n$, $G_w \in \mathbb{R}^{n \times p_w}$ *and* $H_w \in \mathbb{R}^{n_w \times n}, h_w \in \mathbb{R}^{n_w}$. *On the other hand, the state and input constraint sets can be either zonotopes or polytopes, and they are described using the* $\mathcal{H}-$*representation*

$$\mathcal{X} = \{x \in \mathbb{R}^n : H_x x \leq h_x\}, \mathcal{U} = \{u \in \mathbb{R}^m : H_u u \leq h_u\} \tag{4.6}$$

where $H_x \in \mathbb{R}^{n_x \times n}, H_u \in \mathbb{R}^{n_u \times m}, h_x \in \mathbb{R}^{n_x}, h_u \in \mathbb{R}^{n_u}$. *Moreover, the following collection of* $N_t > 0$ *input-state trajectories is available:*

$$\left\{ \left\{ u_k^{(i)} \right\}_{k=0}^{N_s^{(i)}-1}, \left\{ x_k^{(i)} \right\}_{k=0}^{N_s^{(i)}-1} \right\}_{i=1}^{N_t} \tag{4.7}$$

where $N_s^{(i)} > 0$ *is the number of samples in each trajectory. Moreover, the matrix* $\begin{bmatrix} X_-^T & U_-^T \end{bmatrix}^T$ *has full row rank, i.e.,*

$$rank(\begin{bmatrix} X_-^T & U_-^T \end{bmatrix}^T) = n + m \tag{4.8}$$

$$X_- = \begin{bmatrix} x_0^{(1)}, \cdots, x_{N_s^{(1)}-1}^{(1)}, \cdots, x_0^{(N_t)}, \cdots, x_{N_s^{(N_t)}-1}^{(N_t)} \end{bmatrix} \tag{4.9}$$

$$U_- = \begin{bmatrix} u_0^{(1)}, \cdots, u_{N_s^{(1)}-1}^{(1)}, \cdots, u_0^{(N_t)}, \cdots, u_{N_s^{(N_t)}-1}^{(N_t)} \end{bmatrix} \tag{4.10}$$

and $X \in \mathbb{R}^{n \times (N_s+1)N_t}$, $X_- \in \mathbb{R}^{n \times N_s N_t}$, *and* $U_- \in \mathbb{R}^{m \times N_s N_t}$.

The condition (4.8) is fairly standard in the related literature, see, e.g., [44] and references therein. In particular, it ensures that the collected data (4.7) have been obtained for

sufficiently persistent exciting input sequences and that

$$\begin{bmatrix} A & B \end{bmatrix} = X_+ \begin{bmatrix} X_- \\ U_- \end{bmatrix}^\dagger \tag{4.11}$$
$$X_+ = \begin{bmatrix} x_1^{(1)}, \cdots, x_{N_s^{(1)}}^{(1)}, \cdots, x_1^{(N_t)}, \cdots, x_{N_s^{(N_t)}}^{(N_t)} \end{bmatrix}$$

Moreover, the set descriptions (4.5)-(4.6) are similar to the ones made in [48], with the only difference being that we do not restrict the constraints (4.3) to be described by zonotopes.

Problem of interest: *Given the input-state trajectories* (4.7) *collected for the linear model* (4.2)-(4.3) *with unknown system matrices* (A, B) :

(1) Design a data-driven algorithm computing an inner approximation of the ROSC set (4.4)*.*

(2) Design a Data-Driven Set-Theoretic MPC (D-ST-MPC) controller for (4.2)-(4.3) *enjoying the same properties of ST-MPC (see Property 2.1).*

4.3 Proposed Solution

The proposed solution is developed as follows. First, we characterize the set of all system matrices $\mathcal{M}_{\hat{A}\hat{B}}$ consistent with the assumed disturbance bound $\mathcal{Z}_w$ and data (4.7) (see Section 4.3.1). Then, we show how a data-driven zonotopic inner approximation of the ROSC set (2.12) can be obtained from the vertex representation of the set of all compatible system matrices (see Section 4.3.2). Finally, we design a data-driven version of the ST-MPC control strategy by leveraging a family of ROSC sets computed into an extended space domain (see Sections 4.3.3 and 4.3.4).

4.3.1 Data-Driven Representation of Linear Systems With a Bounded Disturbance

In the presence of noise, there generally exist multiple matrices $[A,\ B]$ that are consistent with the data. The following Lemma describes all the linear models consistent with the available data and noise bound $\mathcal{Z}_w$.

Lemma 4.2 *([48, 88]) Let* $T = \sum_{i=1}^{N_t} N_s^{(i)}$ *and consider the following concatenation of multiple noise zonotopes*

$$\mathcal{M}_w = \mathcal{M}_w(C_w, [G_{M_w}^{(1)}, \ldots, G_{M_w}^{(qT)}]),$$

where $C_w \in \mathbb{R}^{n\times(n+m)} = [c_w, \ldots, c_w]$*, and* $G_{M_w} \in \mathbb{R}^{n\times T(n+m)}$ *is built* $\forall\, i \in \{1, \ldots, q\}$, $\forall\, j \in \{2, \ldots, T-1\}$ *as*

$$\begin{aligned} G_{M_w}^{(1+(i-1)T)} &= \begin{bmatrix} g_w^{(i)} & 0_{n\times(T-1)} \end{bmatrix} \\ G_{M_w}^{(j+(i-1)T)} &= \begin{bmatrix} 0_{n\times(j-1)} & g_w^{(i)} & 0_{n\times(T-j)} \end{bmatrix} \\ G_{M_w}^{(T+(i-1)T)} &= \begin{bmatrix} 0_{n\times(T-1)} & g_w^{(i)} \end{bmatrix} \end{aligned} \tag{4.12}$$

Then, the matrix zonotope

$$\begin{aligned} \mathcal{M}_{AB} =& (X_+ - \mathcal{M}_w) \begin{bmatrix} X_- \\ U_- \end{bmatrix}^\dagger \\ :=& \{[\hat{A},\, \hat{B}] : [\hat{A},\, \hat{B}] = C_{AB} + \sum_{i=1}^{T} \beta^{(i)} G_{M_{AB}}^{(i)}, \\ & -1 \le \beta^{(i)} \le 1\} \end{aligned} \tag{4.13}$$

where

$$\begin{aligned} C_{AB} &= (X_+ - C_w) \left([X_-^T,\, U_-^T]^T\right)^\dagger \\ G_{M_{AB}} &= \left[G_{M_w}^{(1)} \left([X_-^T,\, U_-^T]^T\right)^\dagger, \ldots, G_{M_w}^{(qT)} \left([X_-^T,\, U_-^T]^T\right)^\dagger\right] \end{aligned}$$

contains the set of all system matrices $[\hat{A},\, \hat{B}]$ *that are consistent with the data* (4.7) *and disturbance bound* $\mathcal{Z}_w$ *and such that* $[A, B] \in \mathcal{M}_{AB}$*.* □

4.3.2 Data-Driven ROSC Sets

If the system matrices (A, B) are known the ROSC set (4.4) can be computed as [58, Sec. 11.3]:

$$\mathcal{T}^j = \left(\left(\mathcal{T}^{j-1} \ominus \mathcal{Z}_w\right) \oplus \left(-B\mathcal{U}\right)\right) A \cap \mathcal{X} \tag{4.14}$$

In what follows, we aim at computing zonotopic inner approximations of $\mathcal{T}^j$ given the set of all consistent matrices (4.13). To this end, the following issues must be addressed:

(1) The ROSC set $\hat{\mathcal{T}}^j$ consistent with $\mathcal{M}_{AB}$ and $\mathcal{Z}_w$ and such that $\hat{\mathcal{T}}^j \subseteq \mathcal{T}^j$ is:

$$\hat{\mathcal{T}}^j = \left\{ \bigcap_{[\hat{A},\hat{B}]\in\mathcal{M}_{AB}} \left(\left(\hat{\mathcal{T}}^{j-1} \ominus \mathcal{Z}_w \right) \oplus (-\hat{B}\mathcal{U}) \right) \hat{A} \right\} \cap \mathcal{X} \tag{4.15}$$

However, since there are infinite matrices $[\hat{A}, \hat{B}] \in \mathcal{M}_{AB}$, the above is not computable.

(2) According to (4.15), $\hat{\mathcal{T}}^j$ is not (in general) a zonotope even if we assume that $\hat{\mathcal{T}}^{j-1}, \mathcal{U}$ and $\mathcal{X}$ are all zonotopes. Indeed, zonotopes are not closed under Minkowski difference operations and set intersections.

The following proposition is instrumental to obtaining a computable definition of the ROSC set (4.15) and solving the first issue.

Proposition 4.1 *Let $\mathcal{V}_{AB}$ be the set of vertices $\{\hat{A}_i, \hat{B}_i\}_{i=1}^{n_v}$ of the matrix polytope associated to $\mathcal{M}_{AB}$. If $\hat{\mathcal{T}}^{j-1}$ is a convex set, then the ROSC set $\hat{\mathcal{T}}^j$ in (4.15) can be computed as*

$$\hat{\mathcal{T}}^j = \left\{ \bigcap_{[\hat{A}_i,\hat{B}_i]\in\mathcal{V}_{AB}} \left(\left(\hat{\mathcal{T}}^{j-1} \ominus \mathcal{Z}_w \right) \oplus (-\hat{B}_i\mathcal{U}) \right) \hat{A}_i \right\} \cap \mathcal{X} \tag{4.16}$$

Proof 4.1 *Given $\mathcal{V}_{AB}$, we can write any admissible system one-step evolution as $x^+(\alpha) = A(\alpha)x + B(\alpha)u + w$ where*

$$\begin{gathered} \hat{A}(\alpha) = \sum_{i=1}^{n_v} \alpha_i A_i, \quad \hat{B}(\alpha) = \sum_{i=1}^{n_v} \alpha_i B_i \\ [\alpha_1, \ldots, \alpha_{n_v}]^T \in \mathcal{P} = \{\alpha \in \mathbb{R}^{n_v} \colon \alpha_i \geq 0, \sum_{i=1}^{n_v} \alpha_i = 1\} \end{gathered} \tag{4.17}$$

Consequently, the set (4.4) can be written as

$$\begin{aligned} \hat{\mathcal{T}}^j \quad = \quad & \{x \in \mathcal{X} : \exists u \in \mathcal{U} : \sum_{i=1}^{n_v} \alpha_i (\hat{A}_i x + \hat{B}_i u + w) \in \hat{\mathcal{T}}^{j-1}, \\ & \forall w \in \mathcal{Z}_w, \forall \alpha \in \mathcal{P}\} \end{aligned} \tag{4.18}$$

Then, definition (4.18) implies that any convex combination of the vertices must belong to $\hat{\mathcal{T}}^{j-1}$. However, if $\hat{\mathcal{T}}^{j-1}$ is a convex set, then such a requirement is satisfied iff $\forall\, i = 1, \ldots, n_v$, (i.e., for any vertex model $(\hat{A}_i, \hat{B}_i)$) $x_i^+ = \hat{A}_i x + \hat{B}_i u + w \in \hat{\mathcal{T}}^{j-1}$, $\forall w \in \mathcal{Z}_w$.

Consequently, the ROSC set (4.15) *can be computed as in* (4.16)*, concluding the proof.* ■

The second issue can be addressed directly through computation of a polytopic set. However, in this case, the number of vertices necessary to describe the ROSC sets will increase exponentially with the number of computed sets, making such an option not suitable for implementing the ST-MPC controller. On the other hand, a zonotopic inner approximation of the exact ROSC sets can be derived as follows. Consider the $\mathcal{H}-$ representation of $\mathcal{Z}_w$ and $\hat{\mathcal{T}}^{j-1} = \{x \in \mathbb{R}^n : H_{\hat{\mathcal{T}}^{j-1}} x \leq h_{\hat{\mathcal{T}}^{j-1}}\}$, then, we can first compute the polytope $\hat{\mathcal{T}}^j$ using (4.15) and then compute a zonotopic inner approximation of $\hat{\mathcal{T}}^j$, using, e.g., the algorithm proposed in [89, Section IV.A.2] and summarized in Lemma 4.3.

Remark 4.2 *The above solution to compute data-driven ROSC set* $\hat{\mathcal{T}}^j$ *is still not suitable to efficiently implement a data-driven ST-MPC. Indeed, if a family of zonotopic data-driven ROSC sets* $\{\hat{\mathcal{T}}^j\}_{j=0}^N$ *is computed as prescribed by the offline phase of the ST-MPC algorithm (see Section 2.7), then the data-driven equivalent of the optimization* (2.13) *would be*

$$\begin{aligned} u_k = \arg\min_{u \in \mathcal{U}} J(x_k, u) \;\; s.t. \\ \hat{A}_i x + \hat{B}_i u \in (\mathcal{T}^{j_k - 1} \ominus \mathcal{Z}_w),\, \forall i = 1, \ldots, n_v \end{aligned} \tag{4.19}$$

The above presents a number of constraints which scale exponentially with the number of matrix generators describing $\mathcal{M}_{AB}$ *[90]. Consequently, the resulting optimization problem presents a computational complexity much higher than the model-driven counterpart* (2.13)*. In the next section, we solve such an issue by computing an augmented zonotopic representation of the ROSC sets.*

4.3.3 Augmented Data-Driven ROSC Sets

The model-based ROSC set (4.4) can be equivalently re-defined as [26]:

$$\begin{aligned} \Xi^j &= \{(x, u) \in \mathcal{X} \times \mathcal{U} : Ax + Bu + w \in \mathcal{T}^{j-1}, \forall w \in \mathcal{W}\} \\ \mathcal{T}^j &= Proj_x(\Xi^j) \end{aligned} \tag{4.20}$$

where Ξ^j is the $(x, u)-$ augmented space description of the ROSC set and $Proj_x(\Xi^j)$ performs a projection operation of Ξ^j into the $x-$domain. Consequently, for the data-driven

model described by the matrix vertices $\mathcal{V}_{AB}$, we can write

$$\begin{aligned}\hat{\Xi}^j_{AB} = \bigcap_{[\hat{A}_i,\hat{B}_i]\in\mathcal{V}_{AB}} \left\{ z = [x^T, u^T]^T \in \mathbb{R}^{n+m} : H^i_z z \leq h^i_z \right\} \\ \hat{\mathcal{T}}^j = Proj_x(\hat{\Xi}^j_{AB})\end{aligned} \tag{4.21}$$

where

$$H^i_z = \begin{bmatrix} H_x & 0 \\ H_{\hat{\mathcal{T}}^{j-1}}\hat{A}_i & H_{\hat{\mathcal{T}}^{j-1}}\hat{B}_i \\ 0 & H_u \end{bmatrix}, \quad h^i_z = \begin{bmatrix} h_x \\ \tilde{h}_{\hat{\mathcal{T}}^{j-1}} \\ h_u \end{bmatrix} \tag{4.22}$$

and

$$\tilde{h}_{\hat{\mathcal{T}}^{j-1}_r} = \min_{w\in\mathcal{W}} (h_{\hat{\mathcal{T}}^{j-1}_r} - H_{\hat{\mathcal{T}}^{j-1}_r} w) \tag{4.23}$$

Remark 4.3 *For similar reasons to the ones explained in Section 4.3.2, the augmented ROSC set $\hat{\Xi}^j$ is a polytope. Moreover, the projection operation required to obtain $\hat{\mathcal{T}}^j$ is computationally demanding if performed on a polytope [89].*

To obtain a zonotopic inner approximation of $\hat{\Xi}^j$ we can resort to the following lemma:

Lemma 4.3 *(Adapted from [89, Section IV.A.2]) Consider a template zonotope $\mathcal{Z}(c, G)$, with $c \in \mathbb{R}^{n+m}$ denoting any center vector and $G \in \mathbb{R}^{(n+m)\times p}$ a fixed set of chosen generators. The set $\hat{\Xi}^j_z = \mathcal{Z}(c^*, \beta^* G)$ is an inner approximation of $\hat{\Xi}^j$ (i.e., $\hat{\Xi}^j_z \subseteq \hat{\Xi}^j$) if $c^* \in \mathbb{R}^{n+m}, \beta^* =\in \mathbb{R}^{p\times 1}$ are computed solving the following optimization problem*

$$\begin{aligned}\{c^*, \beta^*\} = \arg \max_{c,\, \beta=[\beta_1,\ldots,\beta_l]^T} \sum_{l=1}^{p} d_l \log(\beta_l), \quad s.t. \\ H_{\hat{\Xi}^j} c + |H_{\hat{\Xi}^j} G|\beta \leq h_{\hat{\Xi}^j}, 0 \leq \beta \leq 1\end{aligned} \tag{4.24}$$

where $(H_{\hat{\Xi}^j}, h_{\hat{\Xi}^j})$ denotes the $\mathcal{H}-$representation of $\hat{\Xi}^j$, $d_l \geq 0$ are weighting factors, and $|H_{\hat{\Xi}^j} G|$ denotes a matrix obtained taking the element-wise absolute value of $H_{\hat{\Xi}^j} G$.

□

By denoting with $\mathrm{In}_z(\hat{\Xi}^j)$ the approximation performed by (4.24), the inner zonotopic approximation of (4.21) is

$$\begin{aligned}\hat{\Xi}^j_{AB} &= \bigcap_{[\hat{A}_i,\hat{B}_i]\in\mathcal{V}_{AB}} \left\{z = [x^T, u^T]^T \in \mathbb{R}^{n+m} : H^i_z z \leq h^i_z\right\} \\ \hat{\Xi}^j_z &= \mathrm{In}_z\left\{\hat{\Xi}^j_{AB}\right\}, \quad \hat{\mathcal{T}}^j_z = Proj_x(\hat{\Xi}^j_z)\end{aligned} \tag{4.25}$$

4.3.4 Data-Driven Set-Theoretic MPC

It is now shown how the augmented ROSC sets (4.25) can be used to implement a data-driven ST-MPC strategy, solving the issue discussed in Remark 4.2.

Proposition 4.2 *Given a collection of input-state trajectories satisfying* (4.8)*, then, the Data-Driven Set-Theoretic MPC controller described in Algorithm 4.1 enjoys the same properties as the model-based ST-MPC (see Property 2.1).*

Proof 4.2 *First, in the offline phase, the proposed algorithm computes, using* (4.16)*, a family of pair of ROSC sets* $\{\hat{\Xi}^j_z, \hat{\mathcal{T}}^j_z\}_{j=1}^N$*, where* $\hat{\mathcal{T}}^j_z$ *is an inner zonotopic approximation of* (2.12) *and* $\hat{\Xi}^j_z$ *is an augmented ROSC set representation of* $\hat{\mathcal{T}}^j_z$ *ensuring, by construction, that if* $[x,\, u]^T \in \hat{\Xi}^j_z$ *then* $u \in \mathcal{U}$ *and* $\hat{A}_i x + \hat{B}_i u + w \in \hat{\mathcal{T}}^{j-1}_z, \forall\, w \in \mathcal{W}, \forall \{\hat{A}_i, \hat{B}_i\} \in \mathcal{V}_{AB}$*. Consequently,* $\hat{\Xi}^j_z$ *implicitly embeds the worst-case admissible one-step evolution of the underlying linear system, and the optimization* (4.26) *defines the data-driven equivalent of* (2.13)*. Since by construction* (2.13) *is recursively feasible, then the data-driven MPC Algorithm 4.1 enjoys the same properties as its model-based counterpart.* ■

Algorithm 4.1 Data-Driven Set-Theoretic MPC (D-ST-MPC)

Input: Ver($\mathcal{M}_{AB}$), RCI set $\mathcal{T}^0$, $\mathcal{X}$, $\mathcal{U}$ and $\mathcal{W}$

Offline

1: Let $\hat{\mathcal{T}}_z^0 = \mathcal{T}^0$ a given terminal RCI set

2: Compute the ROSC sets $\{\hat{\Xi}_z^j, \hat{\mathcal{T}}_z^j\}_{i=1}^N$ using (4.25)

Online ($\forall\, k$)

1: Find set membership index $j_k := \min_{j \in \{0,\dots,N\}} \{j : x_k \in \hat{\mathcal{T}}_z^j\}$

2: **if** $j_k = 0$ **then** $u_k = -Kx_k$

3: **else** compute u_k by solving the following QP problem

$$u_k = \arg\min J(x_k, u) \quad s.t. \left[x_k^T, u^T\right]^T \in \hat{\Xi}_z^j \tag{4.26}$$

4: **end if**

Remark 4.4 *The computation of the RCI region $\mathcal{T}^0$ is outside of the scope of this work. However, it can be computed/estimated using data-driven strategies [91]. In alternative, $\mathcal{T}^0$ can be chosen as an arbitrary small set that is only used to initialize the ROSC set computation. With such a solution, the proposed strategy is feasible if $\mathcal{T}^0 \subseteq \hat{\mathcal{T}}_z^1$. Indeed, according to Definition 2.6, if $\mathcal{T}^0 \subseteq \hat{\mathcal{T}}_z^1$, then $\mathcal{T}^0$ is an RCI set and the associated terminal control law for any $x_k \in \mathcal{T}^0$ is*

$$\begin{aligned} u_k = \arg\min J(x_k, u) \quad s.t. \\ \left[x_k^T, u^T\right]^T \in \hat{\Xi}_z^1 \end{aligned} \tag{4.27}$$

4.4 Simulation Results

Consider the discrete-time linear time-invariant system subject to a bounded disturbance described in [59], where

$$A = \begin{bmatrix} 0.7969 & -0.2247 \\ 0.1798 & 0.9767 \end{bmatrix}, \quad B = \begin{bmatrix} 0.1271 \\ 0.0132 \end{bmatrix} \tag{4.28}$$

$|u_k| \leq 3, |x_k| \leq 10$, and the disturbance set is $\mathcal{Z}_w(0_2, 0.005I_2)$. Note that the choice of a two-state system is motivated by the desire to plot the computed data-driven backward reachable sets and better explain the modus operandi of the proposed data-driven ST-MPC. The here shown simulations have been implemented in Matlab R2021a on a Windows 11 pc (16 GB of RAM, I7-11800H - 2.30GHz CPU). The optimizations (2.13)-(4.26) have been solved using SeDuMi and Yalmip, while the CORA toolbox has been used to compute $\mathcal{M}_{AB}$ and $\mathcal{V}_{AB}$. By assuming that the pair (A, B) is unknown, we have simulated the system considering random admissible inputs to collect two input-state trajectories (4.7), each containing $N_s = 10$ samples. After verifying that the rank condition (4.8) is fulfilled, we have computed the set $\mathcal{M}_{AB}$ of all system matrices consistent with (4.7) and $\mathcal{W}$. Then, we computed the vertices $\mathcal{V}_{AB}$ of $\mathcal{M}_{AB}$ and used (4.25) to build a family of $N = 15$ data-driven ROSC sets $\{\hat{\Xi}_z^j, \hat{\mathcal{T}}_z^j\}$ (dashed red zonotopes in Figure 4.1). Since $\hat{\mathcal{T}}_z^1 \subseteq \hat{\mathcal{T}}_z^2$, $\hat{\mathcal{T}}_z^1$ has treated as the terminal RPI set with control law (4.27), see Remark 4.4. To contrast the proposed data-drive ST-MPC with its model-based counterpart, we have also computed a model-based RCI set $\mathcal{T}^0$ using $K = [-0.47, -0.19]$ and the algorithm in [92]. Then, $N = 5$ model-based ROSC sets $\{\mathcal{T}^j\}$ has been computed as prescribed by (2.12) (solid black polytopes in Figure 4.1). Note that the discrepancies in the number of ROSC sets computed for the model-based and data-driven approaches are justified by the fact that the data-driven approach computes inner zonotopic approximations of the actual polytopic sets computed with the model. The results shown in Figures 4.1-4.2 have been obtained for an initial plan's initial state $x_0 = [-2, 1.1]^T$ (i.e., $x_0 \in \hat{\mathcal{T}}_z^{15}$ and $x_0 \in \mathcal{T}^5$) and configuring the ST-MPC and D-ST-MPC schemes to minimize the control effort, i.e., $J(x_k, u) = 0.5u^2$. In Figure 4.1, it is possible to appreciate how both ST-MPC and D-ST-MPC controller are able to confine, in a finite number of steps, the state trajectory in the RCI set $\hat{\mathcal{T}}_z^1$ and $\mathcal{T}_0$, respectively (see the zoom-in in Figure 4.1). This is also confirmed by the set-membership index shown in Figure 4.2 (top subplot). In addition, in Figure 4.2 (bottom subplot), it is possible to note that both controllers produce control input sequences fulfilling the prescribed input constraints. As per the design, D-ST-MPC shows the same behavior as ST-MPC, however, as expected, D-ST-MPC is slightly more conservative in terms of the controller's domain of attraction (the union of the ROSC sets shown in Figure 4.1) and

in terms of time required to reach the terminal RCI set (5 steps for ST-MPC and 8 steps for D-ST-MPC). The average CPU times required to compute the ROSC sets from data and model are 1.61 s and 0.22 s, respectively. On the other hand, the average CPU times to online compute u_k (until $\mathcal{T}^0$ is reached) using D-ST-MPC and ST-MPC are 0.18 s and 0.08 s, respectively.

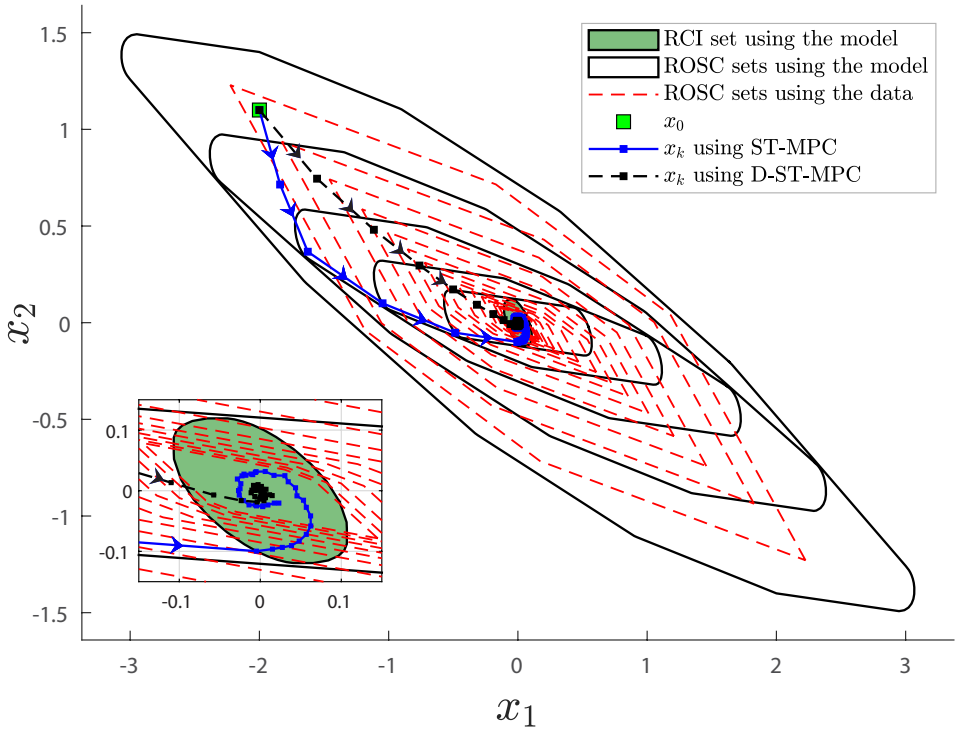


Figure 4.1: Model-based controllable sets $\{\mathcal{T}^j\}_{j=0}^{5}$ (black polyhedral) of system (4.28), projection of augmented data-driven controllable sets $\{\tilde{\Xi}^j\}_{j=0}^{15}$ on x_1 and x_2 axis (red dashed polyhedral)

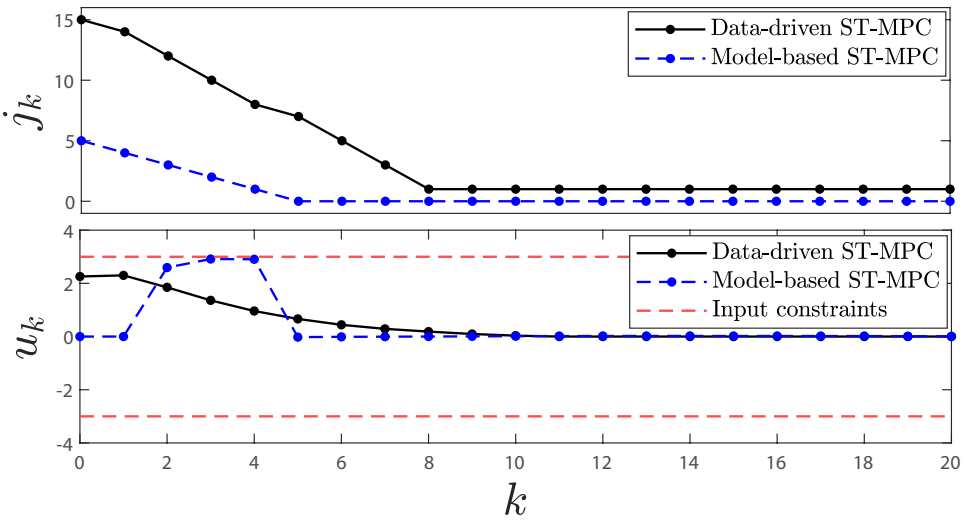


Figure 4.2: Set membership index (top) and control input signals (bottom)

4.5 Conclusions

In this chapter, we proposed a novel data-driven method for computing ROSC sets for an unknown linear system subject to bounded disturbances, state, and input constraints. In particular, it has been shown that using a collection of input-output trajectories, a zonotopic inner approximation of the backward reachable sets can be obtained by resorting to an extended space representation. The peculiar feature of the proposed solution is that the computed sets can also be used to efficiently implement a data-driven dual-mode receding horizon controller. Future studies will be devoted to mitigate the computational complexity and conservativeness of the proposed procedure to compute ROSC sets.

Chapter 5

A Data-Driven Safety Preserving Control Architecture for Constrained Cyber-Physical Systems

5.1 Introduction

In this chapter, we propose a data-driven networked control architecture for unknown and constrained cyber-physical systems capable of detecting networked false-data-injection attacks and ensuring plant's safety. In particular, on the controller's side, we design a novel robust anomaly detector that can discover the presence of network attacks using a data-driven outer approximation of the expected robust one-step reachable set. On the other hand, on the plant's side, we design a data-driven safety verification module, which resorts to worst-case arguments to determine if the received control input is safe for the plant's evolution. Whenever necessary, the same module is in charge of replacing the networked controller with a local data-driven set-theoretic model predictive controller, whose objective is to keep the plant's trajectory in a pre-established safe configuration until an attack-free condition is recovered. Numerical simulations involving a two-tank water system illustrate the features and capabilities of the proposed control architecture.

5.1.1 Contribution of the work

When constrained CPSs are of interest, existing detection and mitigation strategies are developed assuming that an accurate knowledge of the plant's dynamic model is available. Consequently, recent progress made in the field of data-driven control and safety filters, see, e.g., [41–47], have not been fully exploited to design data-driven anomaly detectors or resilient control strategies for constrained CPSs.

The solution proposed in this chapter goes in the direction of filling the identified gap. In particular, it proposes a novel data-driven implementation of the model-based solution presented in [27] for constrained CPSs. The here developed solution leverages the data-driven method developed in [48], which is capable of computing zonotopic outer approximations of forward reachable sets to design a novel and robust data-driven passive anomaly detector local to the controller. Moreover, the data-driven set-theoretic predictive controller developed in [40] is customized to design an emergency controller local to the plant, which is activated whenever the received control input is deemed unsafe according to the expected one-step plant evolution.

The idea behind the proposed emergency controller is, to some extent, related to the concept of safety filters developed in [47] and references therein. In both approaches, a module checks/filters the control inputs generated by a controller to ensure that the plant's state trajectory never leaves a safe set which is explicitly or implicitly defined. However, since in the considered networked setup the reference signal is unavailable on the plant's side, the safety filter would require the use/design of a data-driven terminal robust control invariant region compatible with the tracking domain of the controller. Such a potential drawback is here overcome by designing an emergency controller whose domain of attraction is iteratively enlarged resorting to a family of data-driven robust one-step controllable sets.

A Matlab implementation of the algorithm developed in this chapter is available on the following GitHub page: https://github.com/PreCyseGroup/Data-Driven-Safety-Preserving-Control-Architecture-for-Constrained-CPS.

5.2 Problem Formulation

Consider the class of systems described by a linear time-invariant (LTI) model subject to bounded but unknown disturbance. By denoting with $k \in \mathbb{Z}_+ = \{0, 1, ...\}$ the discrete-time index, and with $x_k \in \mathbb{R}^n$, $u_k \in \mathbb{R}^m$, and $w_k \in \mathbb{R}^n$ the state, control, and disturbance vectors, respectively, the discrete-time LTI evolution is

$$x_{k+1} = Ax_k + Bu_k + w_k \tag{5.1}$$

where A, B are the system matrices of appropriate dimensions and w_k is bounded in a compact set $\mathcal{W} \subset \mathbb{R}^n$. Assume that for physical limitations and safety reasons, the following set-membership constraints must be fulfilled:

$$x_k \in \mathcal{X} \subset \mathbb{R}^n, \quad u_k \in \mathcal{U} \subset \mathbb{R}^m, \tag{5.2}$$

with $\mathcal{X} \subset \mathbb{R}^n, \mathcal{U} \subset \mathbb{R}^m, \mathcal{W}$ zonotopes containing the origin and described using the $\mathcal{G}-$ or $\mathcal{H}-$ representation:

$$\begin{aligned} \mathcal{X} &= \mathcal{Z}_x(c_x, G_x) = \{x \in \mathbb{R}^n : H_x x \leq h_x\} \\ \mathcal{U} &= \mathcal{Z}_u(c_u, G_u) = \{u \in \mathbb{R}^m : H_u u \leq h_u\} \end{aligned} \tag{5.3}$$

$$\mathcal{W} = \mathcal{Z}_w(c_w, G_w) = \{w \in \mathbb{R}^n : H_w w \leq h_w\} \tag{5.4}$$

where $c_x, c_w \in \mathbb{R}^n, c_u \in \mathbb{R}^m$, $G_x \in \mathbb{R}^{n \times p_n}, G_u \in \mathbb{R}^{m \times p_m}, G_w \in \mathbb{R}^{n \times p_w}$, $H_x \in \mathbb{R}^{n_x \times n}, H_u \in \mathbb{R}^{n_u \times m}, H_w \in \mathbb{R}^{n_w \times n}$, $h_x \in \mathbb{R}^{n_x}, h_u \in \mathbb{R}^{n_u}, h_w \in \mathbb{R}^{n_w}$.

Definition 5.1 *At $k \geq 0$, the system (5.1)-(5.2) is said safe if $x_k \in \mathcal{X}$, $u_k \in \mathcal{U}$ and $(Ax_k + Bu_k) \oplus \mathcal{W} \subseteq \mathcal{X}$; unsafe otherwise.*

5.2.1 Networked Setup

Of interest are networked control system setups where the plant and the tracking controller are spatially distributed, and the communication channel is subject to additive False

Data Injection (FDI) attacks. Consequently, the closed-loop evolution of (5.1) is

$$x_{k+1} = Ax_k + Bu_k^{'} + w_k, \qquad u_k = \eta(x_k^{'}, r_k) \tag{5.5}$$

where $r_k \in \mathcal{R} \subset \mathbb{R}^r$ is a bounded reference signal with $\mathcal{R}$ a compact set containing the origin, and $\eta(\cdot,\cdot) : \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}^m$ is the networked tracking controller logic. Moreover, $u_k^{'} := u_k + u_k^a, x_k^{'} := x_k + x_k^a$, with $u_k^a \in \mathbb{R}^m$ and $x_k^a \in \mathbb{R}^n$ the vectors injected by the attacker.

The problem of interest and and the proposed solution are developed under the following assumptions:

Assumption 5.1 *The networked tracking controller $u_k = \eta(x_k^{'}, r_k)$ is given and in the absence of attacks ensures that the plant's constraints* (5.2) *are fulfilled regardless of any admissible disturbance* (5.4) *realization and admissible bounded reference signal. The largest RCI set associated with the tracking controller is hereafter denoted with $\mathcal{X}_\eta \subseteq \mathcal{X}$.*

Assumption 5.2 *The matrices A, B of* (5.1) *are unknown. Moreover, a collection of $N_t > 0$ input-state trajectories is available,*

$$\left\{ \left\{ u_k^{(i)} \right\}_{k=0}^{N_s^{(i)}-1}, \left\{ x_k^{(i)} \right\}_{k=0}^{N_s^{(i)}-1} \right\}_{i=1}^{N_t}, \tag{5.6}$$

where $N_s^{(i)} > 0$ is the number of samples in each trajectory. □

By arranging the collected data into two matrices $X_- \in \mathbb{R}^{n \times N_s N_t}$, $U_- \in \mathbb{R}^{m \times N_s N_t}$, where

$$X_- = \left[x_0^{(1)}, \cdots, x_{N_s^{(1)}-1}^{(1)}, \cdots, x_0^{(N_t)}, \cdots, x_{N_s^{(N_t)}-1}^{(N_t)} \right] \tag{5.7}$$

$$U_- = \left[u_0^{(1)}, \cdots, u_{N_s^{(1)}-1}^{(1)}, \cdots, u_0^{(N_t)}, \cdots, u_{N_s^{(N_t)}-1}^{(N_t)} \right], \tag{5.8}$$

we assume that the matrix $\begin{bmatrix} X_-^T & U_-^T \end{bmatrix}^T$ has full row rank, i.e.,

$$rank(\begin{bmatrix} X_-^T & U_-^T \end{bmatrix}^T) = n + m \tag{5.9}$$

□

Remark 5.1 *Condition* (5.9) *ensures that the collected data* (5.6) *have been obtained for sufficiently persistent exciting input sequences [44].*

In what follows, the one-step ahead shifted representation of X_- is denoted as $X_+ \in \mathbb{R}^{n \times N_s N_t}$, where

$$X_+ = \left[x_1^{(1)}, \cdots, x_{N_s^{(1)}}^{(1)}, \cdots, x_1^{(N_t)}, \cdots, x_{N_s^{(N_t)}}^{(N_t)} \right] \tag{5.10}$$

5.2.2 Problem of Interest

Under the Assumptions 5.1-5.2, design a data-driven networked control architecture such that

- *(O1)* - FDI occurrences are detected before an unsafe condition is reached (see Definition 5.1)
- *(O2)* - The plant remains safe under any FDI attack occurrence, and tracking performance is recovered in the post-attack phase.

5.3 Proposed Data-Driven Control Architecture

The networked data-driven control architecture shown in Figure. 5.1 is hereafter developed, where:

- The *Anomaly Detector* module is a passive device, local to the tracking controller, in charge of detecting anomalies caused by FDI attacks
- The *Safety Verification* module is local to the plant, and it performs safety verification checks on the plant's safety. According to the outcome of these tests, it decides to either use the control signal received by the networked controller or the one provided by the local emergency controller.
- The *Emergency Controller* module is local to the plant, and it aims to confine the state trajectory in a neighbor of an offline defined safe equilibrium point.

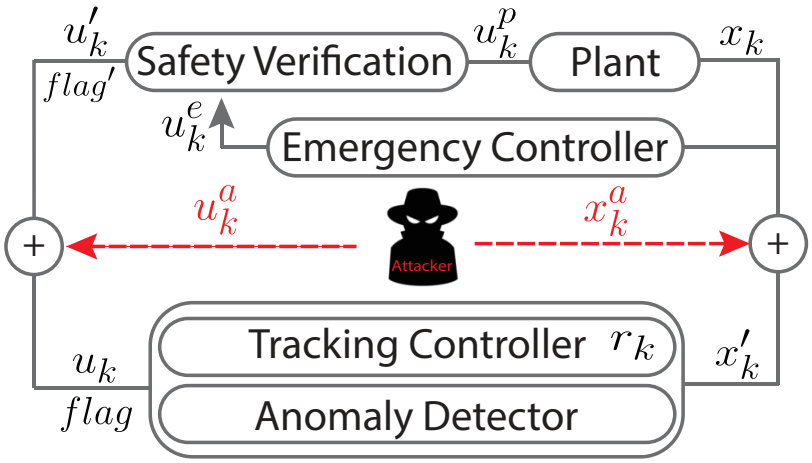


Figure 5.1: Proposed control architecture

Remark 5.2 *It is important to underline that if the attacker launches a coordinated intelligent cover attack on both communication channels, then such an attack cannot be detected by any anomaly detector located on the networked controller's side [21]. This observation motivates the introduction of the emergency controller and a safety verification module shown in Figure. 5.1. Moreover, in different CPS application domains (e.g., Smart Grid), the networked controller is in charge of controlling and coordinating different subsystems that work together to achieve a common goal. In this scenario, only the networked controller is aware of the reference signal that each subsystem must track. Consequently, in the considered setup, the emergency controller cannot replace the networked controller because it is unaware of r_k. Its activation temporarily pauses the tracking task to preserve safety.* □

5.3.1 Anomaly Detector Module

If the model (5.1)-(5.4) is available, then a passive binary anomaly detector can exploit the knowledge of the robust one-step reachable set starting from the current state x_k under u_k for any admissible disturbance realization. Consequently, the detector can claim an anomaly/attack when

$$x'_{k+1} \notin \mathcal{R}^+_k, \quad \mathcal{R}^+_k = Ax'_k \oplus Bu_k \oplus \mathcal{W} \tag{5.11}$$

However, in the considered data-driven setup, the set $\mathcal{R}^+_k$ cannot be directly computed, but it needs to be over-estimated from the available set of data (5.6). The data-driven

computation of an outer approximation of $\mathcal{R}_k^+$ can be obtained by adapting the approach developed in [48, 88]. In particular, first, Lemma 5.1 describes the set of linear models $[\hat{A},\, \hat{B}]$ that are consistent with the data (5.6) and disturbance set $\mathcal{Z}_w$. Then, Lemma 5.2, exploits the results of Lemma 5.1 to build an outer approximation, namely $\hat{\mathcal{R}}_k^+$, of $\mathcal{R}_k^+$.

Lemma 5.1 *[48, Lemma 1] Let* $T = \sum_{i=1}^{N_t} N_s^{(i)}$ *and consider the following concatenation of multiple noise zonotopes*

$$\mathcal{M}_w = \mathcal{M}_w(C_w, [G_{M_w}^{(1)}, \ldots, G_{M_w}^{(qT)}]),$$

where $C_w \in \mathbb{R}^{n\times(n+m)} = [c_w, \ldots,\, c_w]$*, and* $G_{M_w} \in \mathbb{R}^{n\times T(n+m)}$ *is built* $\forall\, i \in \{1, \ldots, q\}$, $\forall\, j \in \{2, \ldots, T-1\}$ *as*

$$\begin{gathered} G_{M_w}^{(1+(i-1)T)} = \begin{bmatrix} g_w^{(i)} & 0_{n\times(T-1)} \end{bmatrix}, \quad G_{M_w}^{(j+(i-1)T)} = \begin{bmatrix} 0_{n\times(j-1)} & g_w^{(i)} & 0_{n\times(T-j)} \end{bmatrix} \\ G_{M_w}^{(T+(i-1)T)} = \begin{bmatrix} 0_{n\times(T-1)} & g_w^{(i)} \end{bmatrix} \end{gathered} \tag{5.12}$$

Then, the matrix zonotope

$$\mathcal{M}_{AB} \;=\; (X_+ - \mathcal{M}_w) \begin{bmatrix} X_- \\ U_- \end{bmatrix}^\dagger := \{[\hat{A},\, \hat{B}] : [\hat{A},\, \hat{B}] = C_{AB} + \sum_{i=1}^{T} \beta^{(i)} G_{M_{AB}}^{(i)}, -1 \le \beta^{(i)} \le 1\} \tag{5.13}$$

where $[\cdot]^\dagger$ *is the right pseudo inverse operator and*

$$\begin{aligned} C_{AB} &= (X_+ - C_w) \left([X_-^T,\, U_-^T]^T\right)^\dagger \\ G_{M_{AB}} &= \left[G_{M_w}^{(1)} \left([X_-^T,\, U_-^T]^T\right)^\dagger, \ldots, G_{M_w}^{(qT)} \left([X_-^T,\, U_-^T]^T\right)^\dagger\right] \end{aligned}$$

contains the set of all system matrices $[\hat{A},\, \hat{B}]$ *consistent with* (5.6) *and* $\mathcal{Z}_w$ *and such that* $[A, B] \in \mathcal{M}_{AB}$. □

Lemma 5.2 *(Adapted from [48, Theorem 1]) The set* $\hat{\mathcal{R}}_k^+ \subset \mathbb{R}^n$*, computed as*

$$\hat{\mathcal{R}}_k^+ = \mathcal{M}_{AB}[x_k', u_k]^T \oplus \mathcal{W} \tag{5.14}$$

is a conservative outer approximation of $\hat{\mathcal{R}}_k^+$.

□

Given, the result of Lemma 5.2, the proposed robust data-driven anomaly detector is (see Figure 5.2)

$$D_k = \begin{cases} \text{anomaly} & \text{if } x'_{k+1} \notin \hat{\mathcal{R}}_k^+ \\ \text{normal} & \text{otherwise} \end{cases} \tag{5.15}$$

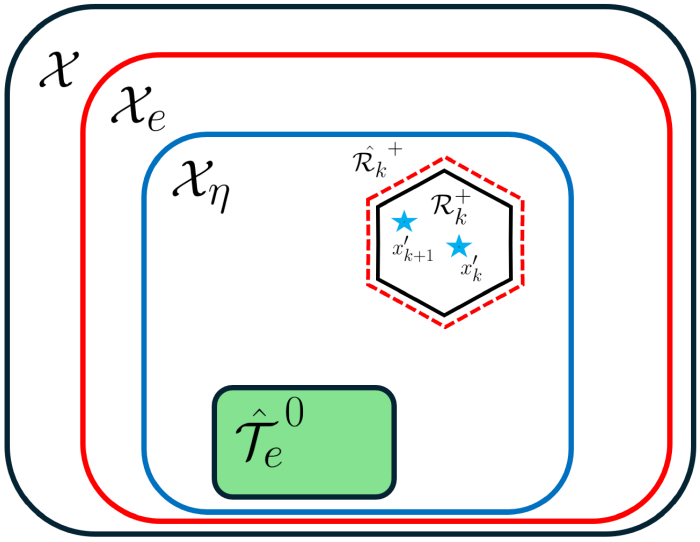


Figure 5.2: Anomaly Detector Logic

Remark 5.3 *Since the detector logic* (5.15) *is based on* (5.14)*, it defines a sufficient conservative condition. Consequently, it ensures the absence of false positives, which is instrumental in avoiding unnecessary activation of the emergency controller with consequent tracking performance losses.* □

In what follows, we assume that when the anomaly detector triggers an anomaly, a $flag = 1$ is transmitted.

5.3.2 Safety Verification Module

The objective of this module is to prevent the plant from reaching unsafe configurations (see Definition 5.1). In particular, given the received u'_k, the safety module checks the following anomalies:

(1) **If** flag ==1 **then** (5.15) detected an attack.

(2) **If** $u'_k \notin \mathcal{U}$ or $\hat{\mathcal{S}}_k^+ \not\subseteq \mathcal{X}_\eta$, with

$$\hat{\mathcal{S}}_k^+ = \mathcal{M}_{AB}[x_k, u'_k]^T \oplus \mathcal{W} \tag{5.16}$$

then the received control input has been corrupted, and it might cause constraints violations.

The condition $\hat{\mathcal{S}}_k^+ \not\subseteq \mathcal{X}_\eta$ ensures that any attack will be detected at least one step before it could compromise the plant's safety. Indeed, if $\hat{\mathcal{S}}_k^+ \not\subseteq \mathcal{X}_\eta$, there is a possibility that x_{k+1} caused by u'_k does not fulfill the constraints (see Figure 5.3). Moreover, in both cases above, the control input u'_k received by the networked controller cannot be applied.

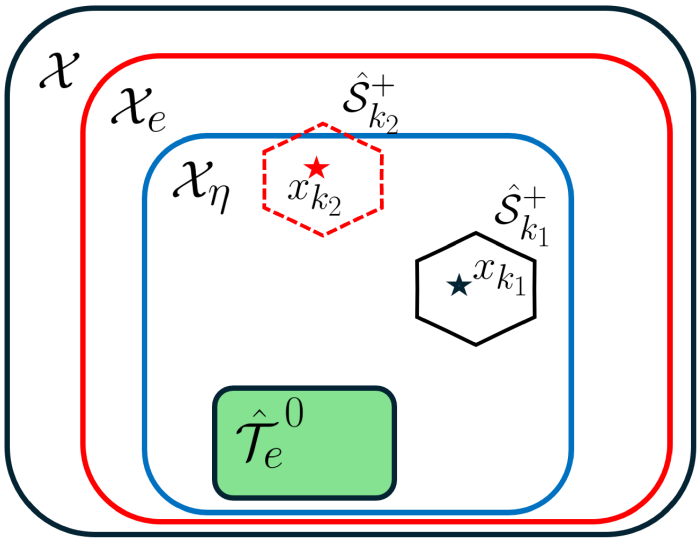


Figure 5.3: Safety Verification Logic

5.3.3 Emergency Controller Module

The objective of the emergency controller is to stabilize the plant around a pre-established equilibrium point whenever the control input received by the networked tracking controller cannot be trusted for safety reasons. By assuming $(0_n, 0_m)$ as the equilibrium pair, the emergency controller

$$u_k^e = f_e(x_k), \quad f_e : \mathcal{X}_e \to \mathcal{U}_e \tag{5.17}$$

must fulfill the following objectives:

(1) $\mathcal{X} \supseteq \mathcal{X}_e \supseteq \mathcal{X}_\eta$ and $\mathcal{U}_e \subseteq \mathcal{U}$;

(2) There exists an RCI set $\hat{\mathcal{T}}_e^0 \subseteq \mathcal{X}_\eta$ where the state trajectory is uniformly ultimately bounded in a finite number of steps

Condition (1) ensures that the emergency controller fulfills all the constraints and that it can be activated anytime and from any state reachable under the tracking controller, while condition (2) guarantees that, in the post-attack phase, the networked controller can be safely re-activated in a finite number of steps.

To meet the objectives *O1-O2*, in what follows, we resort to a customization of the data-driven set-theoretic MPC controller proposed in [40]. The controller's offline design steps can be summarized as follows:

(1) Design a stabilizing data-driven terminal controller, namely $u_k^{e_0} = f_e^0(x_k)$, fulfilling the prescribed constraints and such that the associated RCI set $\hat{\mathcal{T}}_e^0$ is a subset of $\mathcal{X}_\eta$, i.e., $\hat{\mathcal{T}}_e^0 \subseteq \mathcal{X}_\eta$.

(2) If $\hat{\mathcal{T}}_e^0 \subset \mathcal{X}_\eta$, recursively compute a family $\{\hat{\mathcal{T}}_e^j\}_{j=1}^N$ of $N > 0$ sets, with $\hat{\mathcal{T}}_e^j$ an inner approximation of the ROSC set defined in (4.4) and N such that $\bigcup_{j=0}^{N} \hat{\mathcal{T}}_e^j \supseteq \mathcal{X}_\eta$.

The RCI set $\hat{\mathcal{T}}_e^0$ can be computed using the solution in [91, 93] or [40, Remark 5], while $\{\hat{\mathcal{T}}_e^j\}_{j=1}^N$, can be computed using the augmented description proposed in [40, Sec. III.C]. In particular, if the model of the plant is available, the model-based ROSC set defined in Equation (4.4) can be equivalently re-defined as [26]:

$$\begin{aligned} \Xi_e^j &= \{(x,u) \in \mathcal{X} \times \mathcal{U} : Ax + Bu + w \in \mathcal{T}_e^{j-1}, \forall w \in \mathcal{W}\} \\ \mathcal{T}_e^j &= Proj_x(\Xi_e^j) \end{aligned} \tag{5.18}$$

where Ξ_e^j is the $(x,u)-$ augmented space description of the ROSC set and $Proj_x(\Xi_e^j)$ performs a projection operation of Ξ_e^j into the $x-$domain. Moreover, (5.18) can be equivalently

re-written in a matrix form as:

$$\begin{aligned} \Xi_e^j &= \left\{ z = [x^T, u^T]^T \in \mathbb{R}^{n+m} : H^j z \leq h^j \right\} \\ \mathcal{T}_e^j &= Proj_x(\Xi_e^j) \end{aligned} \tag{5.19}$$

with

$$H^j = \begin{bmatrix} H_x & 0 \\ H_{\mathcal{T}_e^{j-1}} A & H_{\mathcal{T}_e^{j-1}} B \\ 0 & H_u \end{bmatrix}, \quad h^j = \begin{bmatrix} h_x \\ \tilde{h}_{\mathcal{T}_e^{j-1}} \\ h_u \end{bmatrix} \tag{5.20}$$

$$[\tilde{h}_{\mathcal{T}_e^{j-1}}]_r = \min_{w \in \mathcal{W}} \left\{ [h_{\mathcal{T}_e^{j-1}}]_r - [H_{\mathcal{T}_e^{j-1}}]_r w \right\} \tag{5.21}$$

and $[h_{\mathcal{T}_e^{j-1}}]_r$, $[H_{\mathcal{T}_e^{j-1}}]_r$ the $r-th$ row of $h_{\mathcal{T}_e^{j-1}}$ and $H_{\mathcal{T}_e^{j-1}}$, respectively. Note that the first row in (5.20) imposes the state constraint $x_k \in \mathcal{X}$, the second and (5.21) that $Ax+Bu+w \in \mathcal{T}_e^{j-1}, \forall w \in \mathcal{W}$, and the third one that $u_k \in \mathcal{U}$. Consequently, for the considered constrained data-driven model, by denoting with $\mathcal{V}_{AB}$ the set of vertices $\hat{A}_i$, $\hat{B}_i \in \mathcal{M}_{AB}$, we can write

$$\begin{aligned} \hat{\mathcal{T}}_e^j = Proj_x(\hat{\Xi}_e^j) = \left\{ x \in \mathbb{R}^n : H_{\hat{\mathcal{T}}_e^j} x \leq h_{\hat{\mathcal{T}}_e^j} \right\}, \\ \hat{\Xi}_e^j = \mathrm{In}_z \left\{ \hat{\Xi}_{AB}^j \right\}, \\ \hat{\Xi}_{AB}^j = \bigcap_{[\hat{A}_i, \hat{B}_i] \in \mathcal{V}_{AB}} \left\{ z = [x^T, u^T]^T \in \mathbb{R}^{n+m} : H_z^i z \leq h_z^i \right\} \end{aligned} \tag{5.22}$$

where $\mathrm{In}_z(\cdot)$ is an operator computing a zonotopic inner approximation of a polytope (e.g., using [89, Sec. IV.A.2]), and

$$H_z^i = \begin{bmatrix} H_x & 0 \\ H_{\hat{\mathcal{T}}_e^{j-1}} \hat{A}_i & H_{\hat{\mathcal{T}}_e^{j-1}} \hat{B}_i \\ 0 & H_u \end{bmatrix}, \quad h_z^i = \begin{bmatrix} h_x \\ \tilde{h}_{\hat{\mathcal{T}}_e^{j-1}} \\ h_u \end{bmatrix} \tag{5.23}$$

with

$$[\tilde{h}_{\hat{\mathcal{T}}_e^{j-1}}]_r = \min_{w \in \mathcal{W}} \left\{ [h_{\hat{\mathcal{T}}_e^{j-1}}]_r - [H_{\hat{\mathcal{T}}_e^{j-1}}]_r w \right\} \tag{5.24}$$

Remark 5.4 *The inner zonotopic approximation of $\hat{\Xi}_{AB}^j$ is needed to efficiently compute the projection $\hat{\mathcal{T}}_e^j$ from $\hat{\Xi}_{AB}^j$, see [40, Remark 4] for a detailed discussion.* □

Given $\{\hat{\mathcal{T}}_e^j\}_{j=0}^N$, and a convex cost function $J(x_k, u)$, the online operations of the Emergency Data-driven Set-Theoretic Controller (E-DSTC) are collected in Algorithm 5.1 (see Figure 5.4).

Algorithm 5.1 E-DSTC - Emergency Controller

Input: RCI set $\hat{\mathcal{T}}_e^0$, and ROSC sets $\{\hat{\Xi}_e^j, \hat{\mathcal{T}}_e^j\}_{i=1}^N$

$(\forall\, k)$

1: Find set membership index $j_k := \min_{j \in \{0,\dots,N\}} \{j : x_k \in \hat{\mathcal{T}}_e^j\}$

2: **if** $j_k = 0$ **then** $u_k^e = f_e^0(x_k)$ **else**

$$u_k^e = \arg\min_u J(x_k, u) \quad s.t. \left[x_k^T, u^T\right]^T \in \hat{\Xi}_e^{j_k} \tag{5.25}$$

3: **end if**

Property 5.1 *For any* $x_0 \in \bigcup_{j=0}^{N} \hat{\mathcal{T}}_e^j$, *the emergency controller E-DSTC ensures that the state trajectory is uniformly ultimately bounded in* $\hat{\mathcal{T}}_e^0$ *in at most* $N-$*steps, i.e.,* $x_k \in \hat{\mathcal{T}}_e^0, \forall k \geq N$ *[40].*

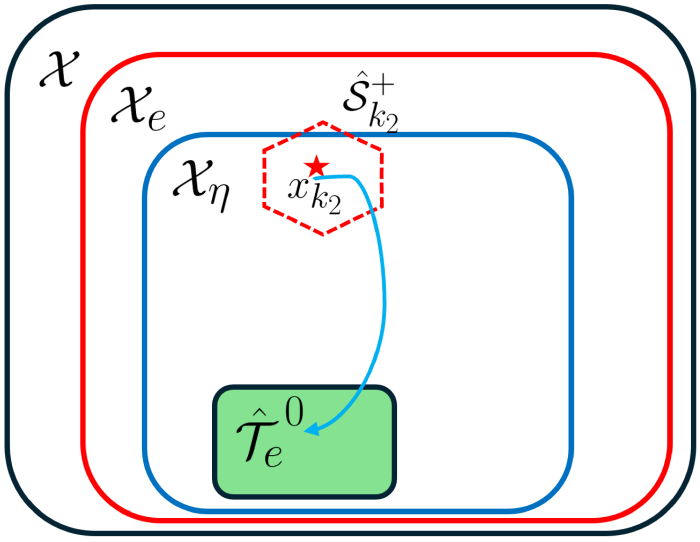


Figure 5.4: Emergency Controller Logic

5.3.4 Switching Control Logic

According to the anomaly test performed by the safety verification module (see subsection 5.3.2), the control input applied to the plant, namely u_k^p, is either u_k' (provided

by the networked controller) or u_k^e (provided by the emergency controller). In this regard, Algorithm 5.2 describes the policy used to switch from u'_k to u_k^e and vice-versa.

Algorithm 5.2 Switching Control Policy

Initialize: emergency = false

$(\forall\, k)$

1: **if** (emergency==true $\wedge\ x_k \in \hat{\mathcal{T}}_e^0$) **then** emergency=false, ignore=1 **else** ignore=0
2: **end if**
3: **if** $(u'_k \notin \mathcal{U}) \vee \left(\hat{\mathcal{S}}_k^+ \not\subseteq \mathcal{X}_\eta\right) \vee$ (flag$'$ ==1 $\wedge$ ignore==0) **then** emergency = true
4: **end if**
5: **if** emergency == true **then** $u_k^p = u_k^e$ **else** $u_k^p = u'_k$
6: **end if**

Proposition 5.1 *The E-DSTC and the Switching Control Policy (Algorithm 5.2) allow the control architecture in Fig. 5.1 to ensure that the plant is safe under any cyber-attacks and performance recovery in the post-attack phase.*

Proof 5.1 *If flag$'$=1 (i.e., an attack is detected by the anomaly detector) is received by the safety verification module, then the condition flag$'$ == 1 used in Step 3 of Algorithm 5.2 triggers the activation of the E-DSTC. On the other hand, if flag$'$=0 (i.e., an attack is not detected by the anomaly detector or flag=1 has been reverted to flag$'$=1 by an attack on the actuation channel), the check* $(u'_k \notin \mathcal{U}) \vee \left(\hat{\mathcal{S}}_k^+ \not\subseteq \mathcal{X}_\eta\right)$ *verifies if the system is safe (see Definition 5.1). If the system is deemed unsafe, then E-DSTC is activated (Step 5). In both cases, since* $\mathcal{X}_\eta \subseteq \bigcup_{j=0}^{N} \hat{\mathcal{T}}_e^j$, *the switch* $u'_k \to u_k^e$ *can be performed fulfilling all the constraints* $\forall\, x \in \mathcal{X}_\eta$. *Once E-DSTC is activated, Step 1 keeps active E-DSTC for at least N step to ensure that* $x_k \in \hat{\mathcal{T}}_e^0 \subseteq \mathcal{X}_\eta$ *(see Property 5.1). When* $x_k \in \hat{\mathcal{T}}_e^0$, *two cases can arise: an attack is still ongoing, or there is a post-attack phase. In both cases, the proposed policy will re-enable the tracking controller if the plant is currently safe under* u'_k, *otherwise, it will keep* $x_{k+1} \in \hat{\mathcal{T}}_e^0$. *Once the tracking controller is restored, Step 3 prescribes to ignore the flag$'$ that might be raised by* (5.15) *at the next control cycle. The latter is motivated by the fact that the current state* x_k *was obtained using* u_{k-1}^e *and not* u_{k-1} *as expected by* (5.15). *Consequently, the proposed control architecture maintains the plant safe under any admissible networked attack, and it ensures tracking performance recovery in the post-attack phase.* □

Remark 5.5 *The data-driven safety preserving architecture designed in Section 5.3 makes use of inner and outer approximations of backward and forward reachable sets. Consequently, the proposed solution is, to some extent, conservative. The conservatism arises from the presence of a bounded unknown disturbance and from the fact that, in a data-driven fashion, we can only compute an over or inner approximation of the actual one-step forward and backward reachable sets. This is explainable from the fact that from the available noisy data, we cannot exactly identify the system matrices* (A, B)*, but we can only compute a set* $\mathcal{M}_{AB}$ *(i.e., a matrix zonotope) of compatible system matrices containing the true matrices* (A, B)*. However, on the other hand, it should also be noted that the sufficient nature of anomaly detector rule* (5.15) *is instrumental in ensuring the absence of false alarm (false positive), which in the proposed architecture avoids unnecessary activation of the emergency module and associated tracking performance losses.*

5.4 Simulation Results

This section shows the effectiveness of the proposed safety-preserving architecture through numerical results obtained considering the two-tank system described in [35]. The system consists of two tanks, which water levels $\bar{h}_i$, $i = 1, 2$ are the state variables, i.e., $\bar{h}_k = [\bar{h}_{1_k}, \bar{h}_{2_k}]^T \in \mathbb{R}^2$. The input vector is $\bar{u} = [\bar{u}_p, \bar{u}_l, \bar{u}_u]^T \in \mathbb{R}^3$, where $\bar{u}_p$ regulates the valve injecting water within the first tank, while $\bar{u}_l$ and $\bar{u}_u$ control the lower and upper valves between the two tanks (see Figure .5.5). By considering the equilibrium point $x_{eq} = [0.5, 0.5]^T$ and $u_{eq} = [0.938, 1, 0.833]^T$ and a sampling time $T_s = 1$, the discrete-time dynamical behaviour of the two-tank system around such a working configuration is described by the following linearized model with a bounded disturbance

$$x_{k+1} = Ax_k + Bu_k + w_k \tag{5.26}$$

where $x_k = \bar{x}_k - x_{eq}$, $u_k = \bar{u}_k - u_{eq}$, $w_k \in \mathcal{W} = \{w \in \mathbb{R}^2 : |w_j| \leq 0.001, j = 1, 2\}$ and

$$A = \begin{bmatrix} 0.993 & 0.003 \\ 0.007 & 0.982 \end{bmatrix}, \quad B = \begin{bmatrix} 0.008 & -0.003 & -0.003 \\ 0.000 & 0.003 & 0.003 \end{bmatrix} \tag{5.27}$$

Moreover, the prescribed state and input constraints are: $-0.7778 \leq u_p \leq 0.6111, -1.25 \leq u_l \leq 0.75, -1.4765 \leq u_u \leq 0.5235$ and $-0.48 \leq h_1, h_2 \leq 0.3$.

However, in the considered data-driven setup, we assume that the matrices A and B are unknown. Consequently, we have simulated the two-tank dynamics around the considered equilibrium point, and by considering random input perturbations, we have collected two input-state trajectories (5.6) which verifies the rank condition (5.9). The collected input-state trajectories and $\mathcal{M}_{AB}$ are available on the provided GitHub page (see Section 5.1.1). From the collected data, using the model identification procedure described in Lemma 5.2, we have obtained the set of all system matrices $\mathcal{M}_{AB}$.

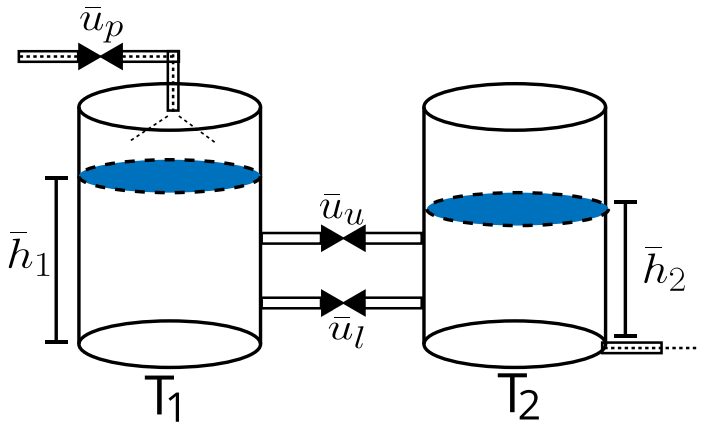


Figure 5.5: Illustration of the considered two-tank system.

In the performed experiments, the networked tracking controller has been designed using the data-driven LQR design procedure proposed in [44], where a saturation module has been added to ensure constraint fulfillment. The E-DSTC controller has been designed using 40 ROSC sets $\{\hat{\mathcal{T}}_e^j\}_{j=1}^{40}$. Moreover, $x_0 = [0.01, -0.01]^T$ and $r_k = [0.1,\ 0.03]^T$ for $k \in [0,\ 100)$ and $r_k = [0.1,\ -0.1]^T$ for $k \in [100,\ 200]$. In the following scenarios, we consider $\mathcal{X}_\eta = \hat{\mathcal{T}}_e^{40}$.

5.4.1 Attack on the Actuation Channel

In the first scenario, the attacker performs an intelligent FDI attack on the actuation channel to affect the plant tracking performance while avoiding triggering the safety checks performed by the safety verification module. In particular, the attacker injects the following

attack vector on the actuation channel

$$u_k^a = \begin{cases} \arg\min_{u^a} \|u^a - [-1,\ -2,\ -2]^T\|_2^2, \ \text{s.t. } u^a + u_k \in \mathcal{U}, & \text{if } k \in [40,\ 60] \\ \arg\min_{u^a} \|u^a - [2,\ -2,\ 1]^T\|_2^2, \ \text{s.t. } u^a + u_k \in \mathcal{U}, & \text{if } k \in [145,\ 154] \\ 0, & \text{Otherwise} \end{cases} \tag{5.28}$$

and it sets $flag = 0$ when $u_k^a \neq 0$. The obtained results are shown in Figure. 5.6 - Figure. 5.9. When the attack starts at $k = 40$, the state trajectory deviates from the desired reference (see the red line in Figure. 5.6). At $k = 41$, the anomaly detector (5.15) detects an anomaly (see Figure. 5.7) because the received measurement $x'_{41} \notin \mathcal{R}^+_{40}$. However, since the attacker corrupts the anomaly flag, then an emergency is not raised by the safety checks until $k = 52$ when the plant is one-step away from possibly violating the constraints, i.e., $\hat{\mathcal{S}}^+_{52} \not\subseteq \hat{\mathcal{T}}_e^{40}$ (see the bottom left small box in Figure. 5.6). Given the safety risk, at $k = 52$, the emergency controller is activated to steer the state of the system into the safety region $\hat{\mathcal{T}}_e^0$ (the green polyhedral set in Figure. 5.6) without constraint violation (see the green line in Figure. 5.6).

When at $k = 62$ the state trajectory reaches $\hat{\mathcal{T}}_e^0$, the networked tracking controller can be safely restored since the attack is terminated. Consequently, the plant can recover and track the desired r_1 (see the magenta dashed line in Figure. 5.6). At $k = 145$, when the attacker launches the second attack vector on the actuation channel, the safety verification module detects the presence of the attack at $k = 145$ since the received control input, $u_k^{'}$ is not one-step safe, i.e., $\hat{\mathcal{S}}^+_{145} \not\subseteq \hat{\mathcal{T}}_e^{40}$ (see the top right small box in Figure. 5.6). Hence, at $k = 145$, the emergency controller is activated and steers the state of the system into $\hat{\mathcal{T}}_e^0$. When at $k = 151$ the state trajectory reaches $\hat{\mathcal{T}}_e^0$, the emergency controller is deactivated. However, since the attacker action has not been terminated until $k = 154$ (see the red line in Figure. 5.6), the state trajectory is still under the influence of the attacker's injection without compromising the safety of the plant. Finally, at $k = 155$, the networked tracking controller is safely restored and steers the state of the system to r_{100}. The control signals applied to the plant are shown in Figure. 5.8, which show that using the proposed control strategies, the input constraints are always fulfilled.

To better show the effectiveness of the combined use of the safety verification module

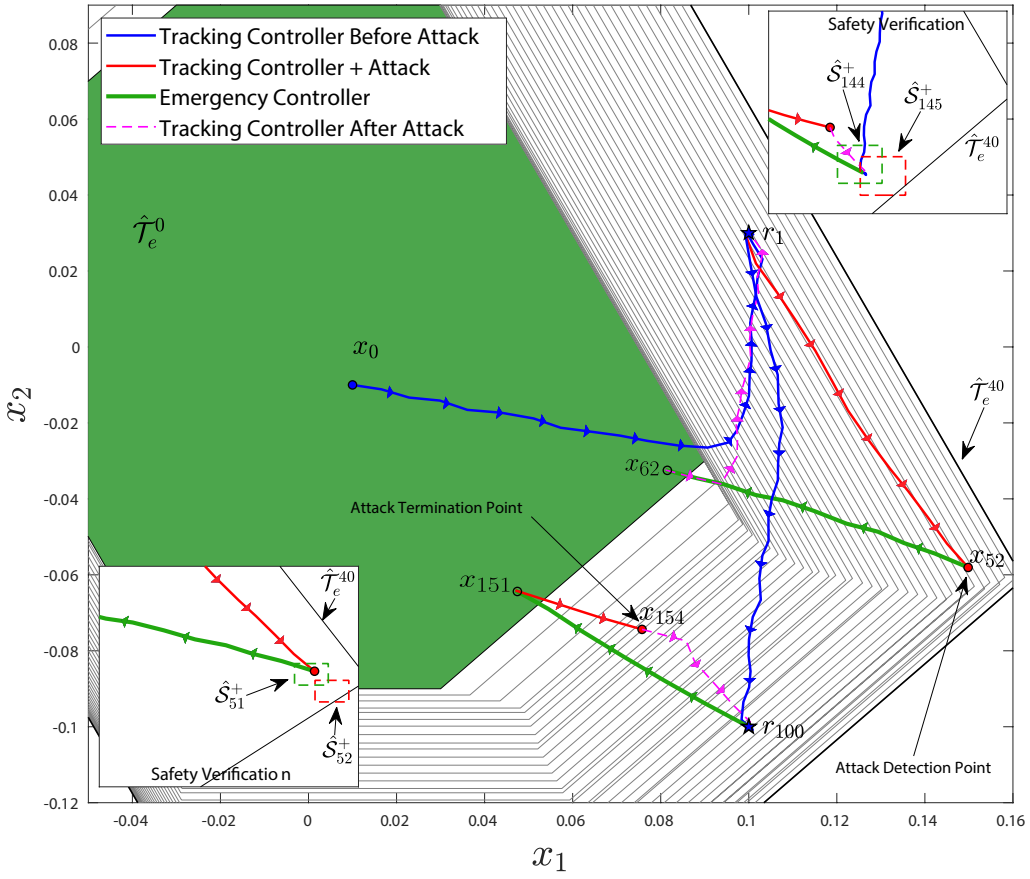


Figure 5.6: Case A: State Evolution and Emergency Controller Operation

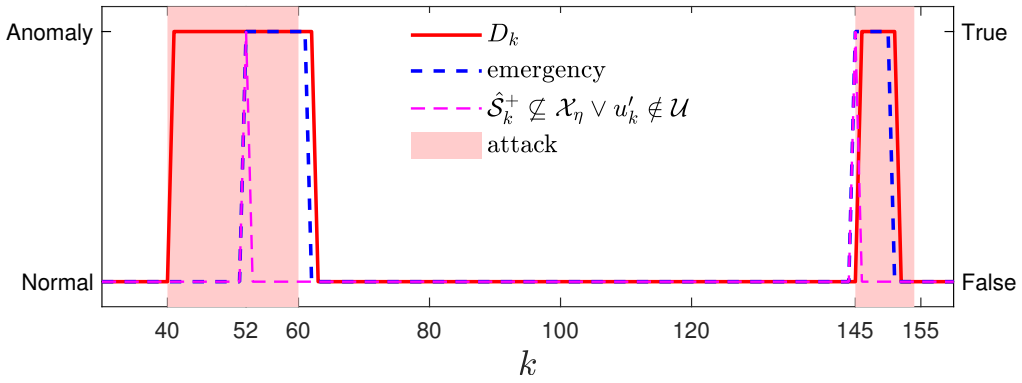


Figure 5.7: Case A: Detector and Safety Verification Outputs

and emergency controller, the proposed data-driven control architecture has been contrasted with the same scheme where the safety features (safety verification and emergency controller module) have been disabled, see Figure. 5.9. The results depicted in Figure. 5.9 show that in the absence of safety features on the plant's side, the attacker is able to steer the plant's state trajectory outside of the safe region $\mathcal{X}_\eta$, i.e., $x_k \notin \hat{\mathcal{T}}_{40}$ (see the magenta dashed line on

the right subplot of Figure. 5.9). Moreover, Table 5.1 numerically quantifies the advantages of the proposed solution by reporting the % of safety constraints violations ($\%violations$) occurring over the total simulation time for both the proposed approach and scheme without safety modules, where

$$\%violations = \frac{\sum_{i=1}^{T_{sim}} \left\{1 \textbf{ if } x(k) \notin \hat{\mathcal{T}}_e^{40}, 0 \textbf{ otherwise}\right\}}{T_{sim}}\%$$

The results show that in the absence of our safety mechanism, the attacker is able to steer the plant state trajectory outside of the safety region for 24% of the total simulation time.

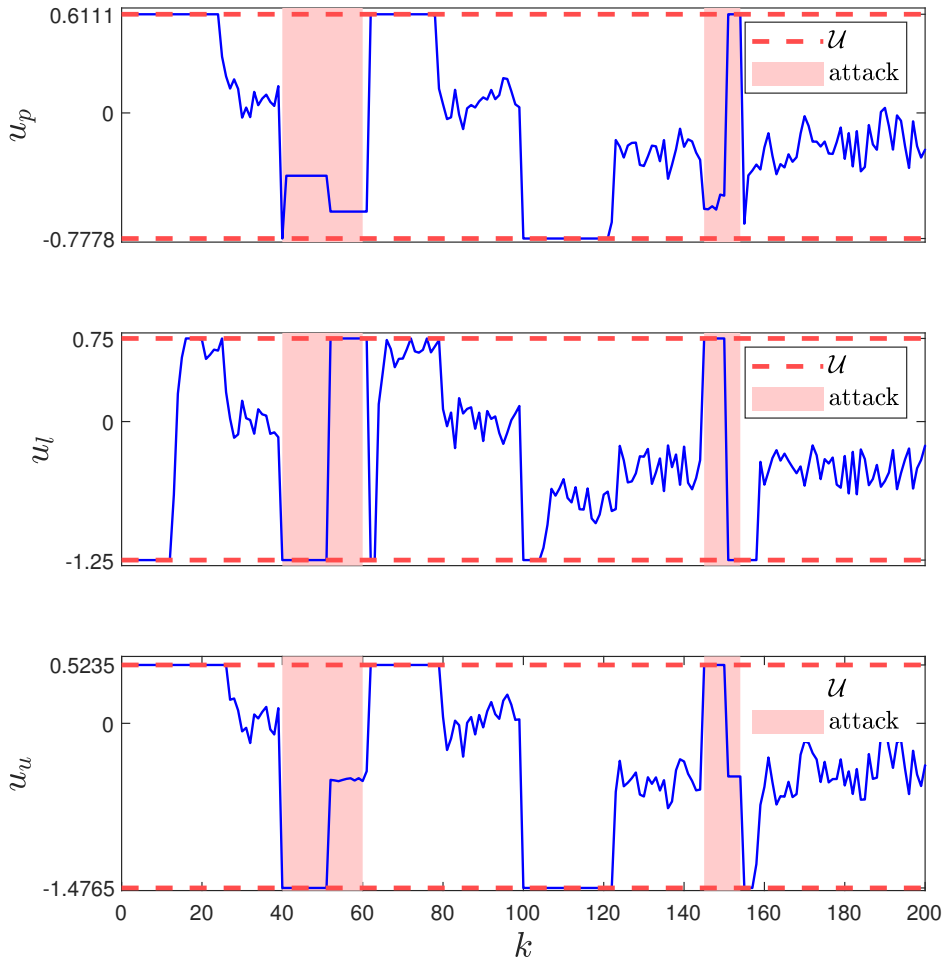


Figure 5.8: Case A: Control Signals

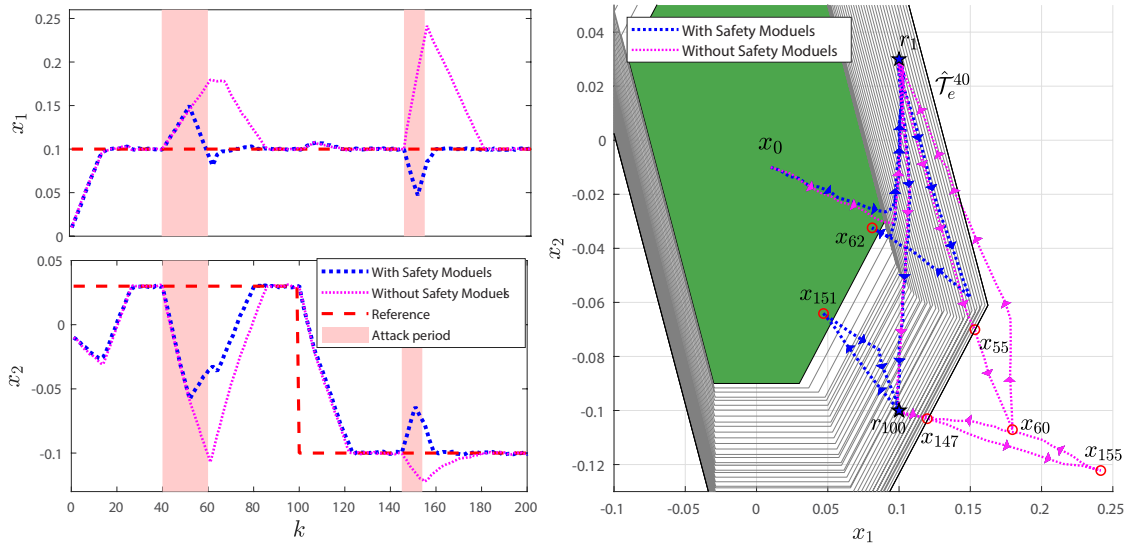


Figure 5.9: Case A: proposed data-driven control architecture vs control scheme without the safety modules.

Table 5.1: Case A: % of safety constraints violations over the total simulation time: proposed approach vs control scheme without safety modules

	Control scheme without safety modules	Proposed approach
%violations	24%	0%

5.4.2 Attack on the Measurement Channel

In the second scenario, the attacker performs an FDI on the measurement channel. In particular, for $k \in [95, 113]$, it corrupts x_k by adding on it the vector $[0.0025(k-95), 0]^T$. The obtained results are shown in Figures 5.10-5.13. Since at the beginning of the attack period, the magnitude of the injection is minimal and comparable with the one of the process disturbance $\mathcal{W}$, the attack is undetected by (5.15) until $k = 109$ (see Figure. 5.11). At $k = 110$, the attack is detected by (5.15) because $x'_{110} \notin \hat{\mathcal{R}}^+_{109}$ (see small box in Figure. 5.10). Consequently, the anomaly detector transmits flag=1 to the plant, triggering the activation of the emergency controller (see green line in Figure. 5.10). At $k = 115$, under the E-DSTC, the system's state reaches $\hat{\mathcal{T}}_e^0$ while the attack is terminated. Consequently, at $k = 115$, the anomaly still raised by (5.15) is ignored for one step (see Step (3) of Algorithm. 5.2) and the tracking controller is reactivated. For $k \geq 115$, the state trajectory resumes tracking r_k (see the magenta dashed line in Figure. 5.10).

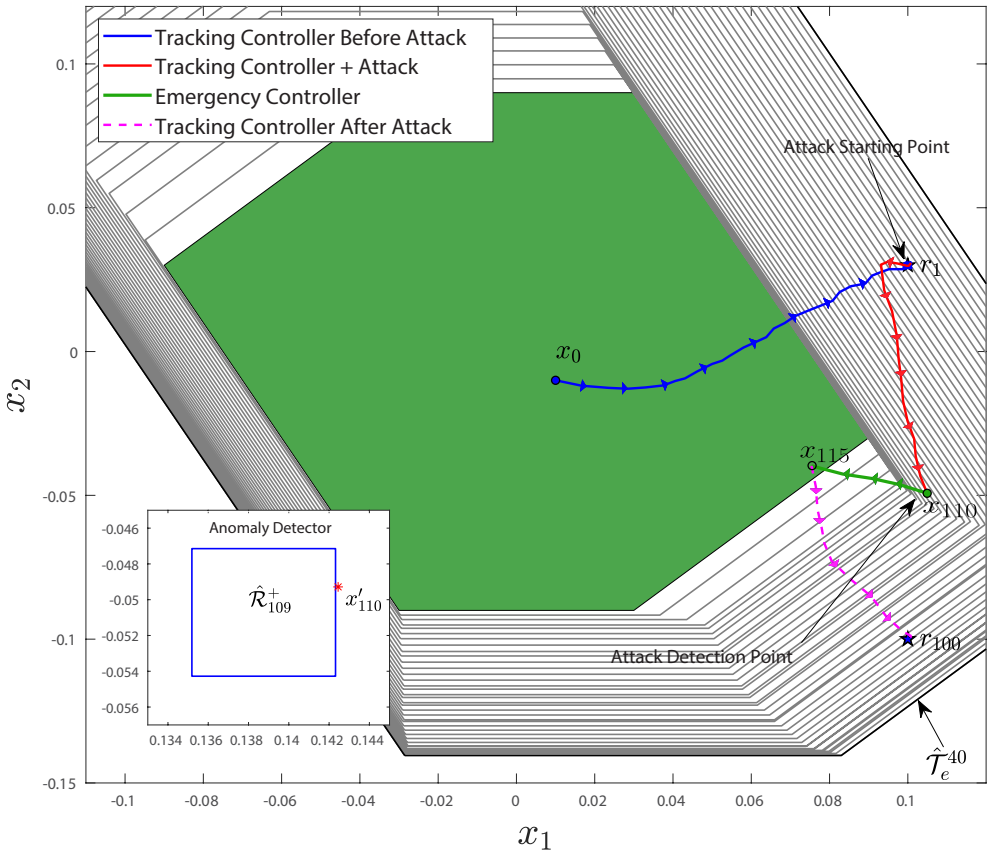


Figure 5.10: Case B: State Evolution and Detectors Operation

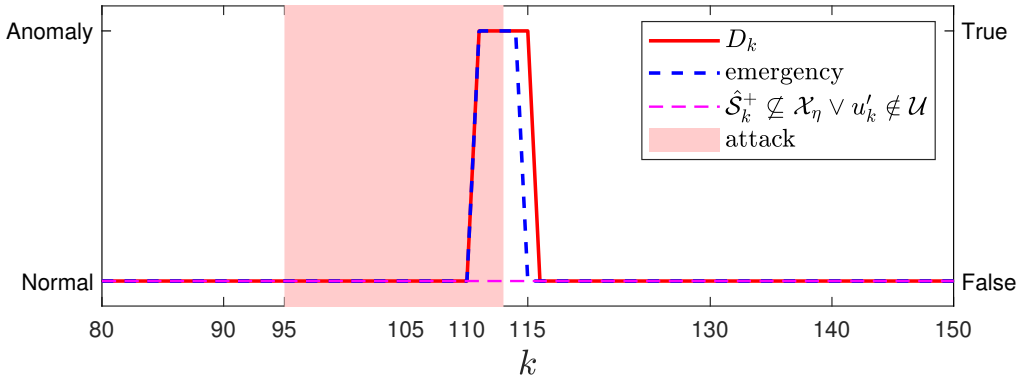


Figure 5.11: Case B: Detector and Safety Verification Outputs

On the other hand, the control signals applied to the plant presented in Figure.5.12 show that the proposed control strategies always fulfill the prescribed input constraint set.

As in the previous scenario, the proposed data-driven control architecture has been contrasted with the same scheme where the safety features (safety verification and emergency

controller module) have been disabled. The results are displayed in Figure.5.13. Specifically, from $k = 101$, until $k = 173$, in the absence of the safety features, the attacker is able to steer the plant's state trajectory outside safety region, i.e., $x_k \notin \hat{\mathcal{T}}_e^{40}$ $k \in [101, 173]$ (see the magenta dashed line on the right subplot of Figure. 5.13). Finally, Table 5.2 shows that in the absence of our safety mechanism the attacker is able to steer the plant state trajectory outside of the safety region for 36% of the total simulation time.

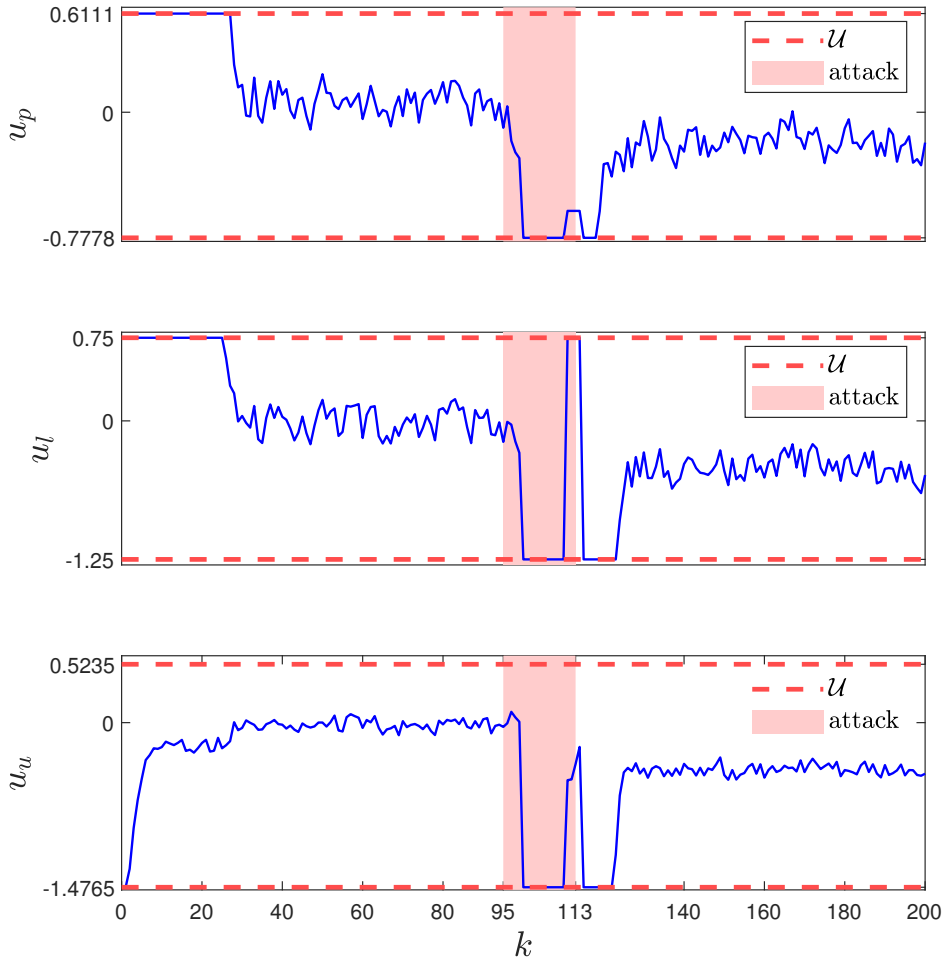


Figure 5.12: Case B: control signals

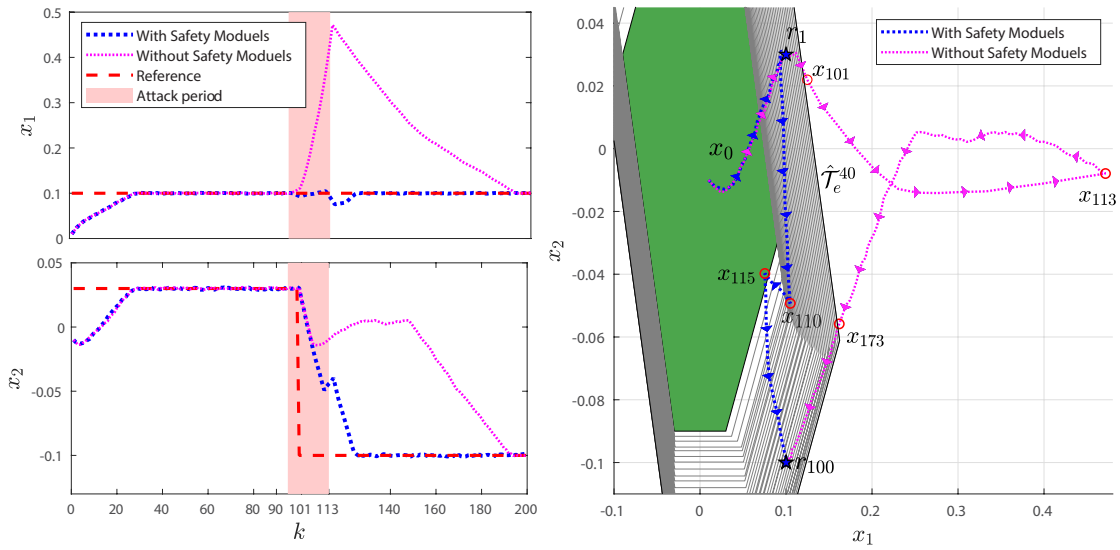


Figure 5.13: Case B: proposed data-driven control architecture vs control scheme without the safety modules.

Table 5.2: Case B: % of safety constraints violations over the total simulation time: proposed approach vs control scheme without safety modules

	Control scheme without safety modules	Proposed approach
%violations	36%	0%

5.5 Conclusions

In this chapter, a data-driven safety-preserving control architecture for constrained CPS has been proposed. The proposed solution leveraged worst-case backward and forward reachability arguments to develop a data-driven robust anomaly detector on the networked controller and a safety-preserving module on the plant. Moreover, a safe switching mechanism between the networked and local emergency controllers has been defined to ensure safety during the attack and performance recovery thereafter. It has been proved that the proposed architecture ensures attack detection at least one step before the plant's safety could be compromised. Simulation results have been shown to illustrate the capability of the proposed data-driven control architecture. Future works will be devoted to reducing the conservativeness of the proposed design, extending it to a more general nonlinear plant model, and mitigating the tracking performance degradations when the emergency controller is activated.

Chapter 6

A Data-Driven Approach To Preserve Safety and Reference Tracking for Constrained Cyber-Physical Systems Under Network Attacks

6.1 Introduction

In this chapter, a worst-case data-driven control architecture capable of ensuring the safety of constrained Cyber-Physical Systems under cyber-attacks while minimizing, whenever possible, potential degradation in tracking performance is proposed. To this end, a data-driven robust anomaly detector is designed to detect cyber-attack occurrences. Moreover, an add-on tracking supervisor module allows safe open-loop tracking control operations in case of unreliable measurements. On the plant side, a safety verification module and a local emergency controller are designed to manage severe attack scenarios that cannot be handled on the controller's side. These two modules resort to worst-case reachability and controllability data-driven arguments to detect potential unsafe scenarios and replace, whenever strictly needed, the tracking controller with emergency actions whose objective is to steer the plant's state trajectory in a predefined set of admissible and safe robust control invariant region until an attack-free scenario is restored. The effectiveness of the proposed

solution has been shown through a simulation example.

6.1.1 Contribution of the Work

Based on the reviewed literature in Chapter 1, most of the existing results on the data-driven design of safety preserving control architectures neglect the tracking performance degradation problem under cyber-attacks. In particular, to the best of the author's knowledge, the problem of preserving the safety and tracking performance of a constrained CPS from a collection of noisy input-state trajectories has not yet been explored in the literature. Consequently, this chapter goes in the direction of filling the existing gap, developing a solution that tries to minimize, whenever possible and in a data-driven fashion, the tracking performance loss under cyber-attacks. In particular, we here extend the model-based solution in [35] and [2] to develop a novel data-driven control architecture that is capable of preserving the safety of the plant while minimizing, whenever possible, the tracking performance loss due to cyber-attacks on the communication channels. The main contributions of this work can be summarized as follows:

- Unlike the approach in [27], the proposed solution is entirely data-driven.
- To the best of the author's knowledge, existing data-driven solutions only focus on preserving the safety of the plant (see, e.g., [2, 47, 94]). On the other hand, the here proposed solution ensures safety while it minimizes, whenever possible, the tracking performance loss during attack scenarios.
- Differently from [35], we here do not assume that attacks can be detected without delay (i.e., by using authenticated communication channels). Consequently, a more general setup that considers attack detection delays is developed.

A Matlab implementation of the algorithm developed in this chapter is available on the following GitHub page: https://github.com/PreCyseGroup/data-driven-tracking-supervisor.

6.2 Preliminaries and Problem Formulation

Denote with $k \in \mathbb{Z}_+ = \{0, 1, \ldots\}$ a discrete-time index, and consider the discrete-time Linear Time-Invariant (LTI) system

$$z_{k+1} = \Phi z_k + G\mu_k + p_k, \quad z_k \in \mathcal{Z},\ \mu_k \in \mathcal{U}_\mu,\ p_k \in \mathcal{P} \tag{6.1}$$

where p_k is an unknown but bounded process disturbance and $\mathcal{Z}$, $\mathcal{U}_\mu$, and $\mathcal{P}$ are compact sets.

Definition 6.1 *Consider the LTI system* (6.1)*, the Robust One-Step Reachable Set (RORS)* $\mathcal{R}$ *is defined as follows [58, Section 11.3]:*

$$\mathcal{R} = \{z^+ : \exists z \in \mathcal{Z}, \mu \in \mathcal{U}_\mu, w \in \mathcal{W}\ s.t.\ z^+ = \Phi z + G\mu + p\} \tag{6.2}$$

Definition 6.2 *At* $k \geq 0$*, the system* $z_{k+1} = \Phi z_k + G\mu_k + p_k$*, subject to constraints is said safe if* $z_k \in \mathcal{Z}$, $\mu_k \in \mathcal{U}_\mu$ *and* $(\Phi z_k + G\mu_k) \oplus \mathcal{P} \subseteq \mathcal{Z}$*; unsafe otherwise.*

Of interest for this chapter are networked control systems setups where the plant and the tracking controller are spatially distributed and a communication medium is used to exchange state measurements and control inputs. By considering the possibility of cyber-attacks on the communication channels, we also assume that an anomaly detector is implemented local to the controller. In particular, the assumed plant model, controller, cyber-attack actions, and anomaly detector logic can be formalized as follows.

6.2.1 Plant Model and Safety Constraints

We consider plants whose dynamic evolution can be described by means of the following LTI system

$$x_{k+1} = Ax_k + Bu_k + w_k \tag{6.3}$$

where $x_k \in \mathbb{R}^n, u_k \in \mathbb{R}^m$ and A, B are the unknown system matrices and w_k is a bounded disturbance that lies into a known compact set $\mathcal{W} \subset \mathbb{R}^n$. Due to physical limitations and

safety reasons, the following set-membership constraints are prescribed:

$$x_k \in \mathcal{X} \subset \mathbb{R}^n, \quad u_k \in \mathcal{U} \subset \mathbb{R}^m, \tag{6.4}$$

where we assume that $\mathcal{X} \subset \mathbb{R}^n, \mathcal{U} \subset \mathbb{R}^m, \mathcal{W} \subset \mathbb{R}^n$ are compact sets described by means of Zonotopes containing the origin.

6.2.2 Tracking Controller

The plant is required to track a reference signal r_k where $r_k \in \mathcal{R} \subset \mathbb{R}^r$ is a bounded reference signal with $\mathcal{R}$ a compact set containing the origin. Moreover, it is assumed that a networked stabilizing tracking controller is available with the following model:

$$u_k = \eta(x_k, r_k) \tag{6.5}$$

where $\eta(\cdot, \cdot) : \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}^m$ is the networked tracking controller logic.

Assumption 6.1 *The networked tracking controller* (6.5) *is given, and in the absence of attacks, it ensures that the plant's constraints* (6.4) *are satisfied regardless of any realization of the admissible disturbance $\mathcal{W}$ and any admissible bounded reference signal. The largest RCI set associated with the tracking controller is hereafter denoted by $\mathcal{X}_\eta \subseteq \mathcal{X}$.*

6.2.3 Networked Cyber-Attacks

We assume that the communication channels between the plant and the controller are vulnerable to False Data Injection (FDI) attacks. Consequently, the closed-loop evolution of (6.3) under FDI attacks on both the actuation and measurement channel is

$$x_{k+1} = Ax_k + Bu'_k + w_k, \quad u_k = \eta(x'_k, r_k) \tag{6.6}$$

where $u'_k := u_k + u^a_k, x'_k := x_k + x^a_k$, with $u^a_k \in \mathbb{R}^m$ and $x^a_k \in \mathbb{R}^n$ the vectors injected by the attacker.

6.2.4 Anomaly Detector

We assume that a passive data-driven binary anomaly detector is available on the controller's side to detect the presence of cyber-attacks by leveraging the received state measurements $\{x'_t\}_{t=0}^{k}$ and computed control inputs $\{u_t\}_{t=0}^{k}$. The following model abstractly describes the anomaly detector

$$d_k = D\left(\{x'_t\}_{t=0}^{k}, \{u_t\}_{t=0}^{k-1}, \mathcal{W}\right) \tag{6.7}$$

where $D(\cdot,\cdot,\cdot)$ is the detection logic and $d_k \in [0,1]$. In what follows, we will assume that $d_k = 1$ denotes the presence of an anomaly. Moreover, it is assumed that the detection mechanism is capable of providing attack detection with an estimated bounded delay $0 \leq \tau < \infty$ [95, 96].

Assumption 6.2 *The matrices A, B of* (6.3) *are unknown. Moreover, a collection of $N_t > 0$ input-state trajectories is available,*

$$\left\{ \left\{u_k^{(i)}\right\}_{k=0}^{N_s^{(i)}-1}, \left\{x_k^{(i)}\right\}_{k=0}^{N_s^{(i)}-1} \right\}_{i=1}^{N_t}, \tag{6.8}$$

where $N_s^{(i)} > 0$ is the number of samples in each trajectory. By arranging the collected data into two matrices $X_- \in \mathbb{R}^{n\times N_s N_t}$, $U_- \in \mathbb{R}^{m\times N_s N_t}$, where

$$X_- = \left[x_0^{(1)}, \cdots, x_{N_s^{(1)}-1}^{(1)}, \cdots, x_0^{(N_t)}, \cdots, x_{N_s^{(N_t)}-1}^{(N_t)}\right] \tag{6.9}$$

$$U_- = \left[u_0^{(1)}, \cdots, u_{N_s^{(1)}-1}^{(1)}, \cdots, u_0^{(N_t)}, \cdots, u_{N_s^{(N_t)}-1}^{(N_t)}\right], \tag{6.10}$$

we assume that the matrix $\begin{bmatrix} X_-^T & U_-^T \end{bmatrix}^T$ has full row rank, i.e.,

$$rank(\begin{bmatrix} X_-^T & U_-^T \end{bmatrix}^T) = n + m \tag{6.11}$$

□

Remark 6.1 *Condition* (6.11) *ensures that the collected data* (6.8) *have been obtained for*

sufficiently persistent exciting input sequences [44]. In what follows, the one-step ahead shifted representation of X_- is denoted as $X_+ \in \mathbb{R}^{n\times N_s N_t}$, where

$$X_+ = \left[x_1^{(1)}, \cdots, x_{N_s^{(1)}}^{(1)}, \cdots, x_1^{(N_t)}, \cdots, x_{N_s^{(N_t)}}^{(N_t)} \right] \tag{6.12}$$

6.2.5 Problem Formulation

In the considered networked architecture, we assume that only the networked controller is aware of the reference signal r_k. Therefore, it is acceptable to experience performance degradation during cyber-attacks as long as the plant's safety and recovery (after the attack) are guaranteed. However, it is also desirable to design the control architecture to minimize the tracking performance degradation due to cyber-attacks. Consequently, the problem of interest can be stated as follows:

Under Assumptions 6.1-6.2, design a data-driven control architecture for the constrained system (6.3)- (6.4) ensuring

(1) Plant's safety and tracking performance loss minimization during any cyber-attack, and

(2) Performance recovery in the post-attack phase.

6.3 Proposed Solution

To design the proposed data-driven solution, we consider a worst-case scenario in which a cyber-attack can affect both communication channels, either simultaneously or at different times. If only safety is of interest, then a data-driven solution to the problem can be obtained by adapting the strategy developed in [2]. Indeed, upon detecting an attack, such a strategy prescribes disconnecting the networked tracking controller and activating a local emergency controller for safety-preserving purposes. The drawback of such a solution is that it jeopardizes the tracking task regardless of the nature of the attack, producing an overly conservative mitigation strategy that ultimately results in poor tracking performance. On the other hand, in the here proposed solution, we extend the approach in [2] with

the aim of minimizing, whenever possible, the tracking performance loss caused by cyber-attacks. In particular, we argue that tracking performance degradation can be minimized by implementing targeted measures to counteract attacks on the actuation and measurement channels. To this end, we enhance the networked control scheme shown in Figure. 6.1 with four additional modules:

- A Safety Verification (SV) subsystem, local to the plant, whose objective is to prevent the plant from reaching unsafe configurations (see Definition 6.2);
- An Emergency Controller (EC), local to the plant, in charge of guaranteeing the plant safety when the received control signal $u_k^{'}$ is untrustworthy.
- An Anomaly Detector (D) in charge of detecting the presence of attacks.
- A Tracking Supervisor (TS), local to the tracking controller, responsible for minimizing the tracking performance loss while ensuring safety when $x_k^{'}$ is invalid. This module is activated when the Anomaly Detector detects the presence of a cyber-attack (i.e., if $d_k = 1$).

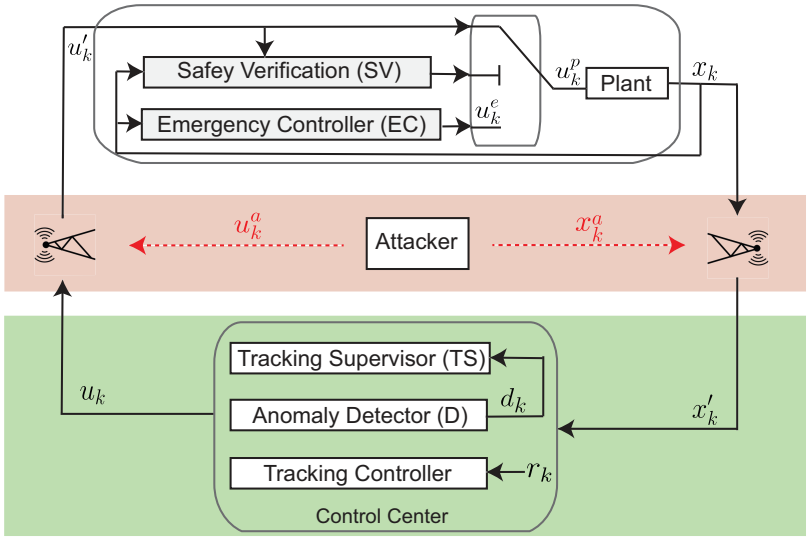


Figure 6.1: Proposed Control Architecture

6.3.1 Safety Verification Module

This module aims to prevent the plant from reaching unsafe configurations (see Definition 6.2). In particular, given the received u'_k, the safety module checks the following possible anomalies:

$$u'_k \notin \mathcal{U}, \qquad \mathcal{S}_k^+ \not\subseteq \mathcal{X}_\eta \tag{6.13}$$

where $\mathcal{S}_k^+$ denotes the RORS set starting from x_k and u'_k. Therefore, $\mathcal{S}_k^+ \not\subseteq \mathcal{X}_\eta$ represents a situation where exists a disturbance realization such that x_{k+1} does not fulfill the constraints. Consequently, the safety verification logic can be summarized as follows

- **If** $u'_k \notin \mathcal{U}$ or $\mathcal{S}_k^+ \not\subseteq \mathcal{X}_\eta$, **then** the u'_k has been corrupted, and the *EC* is activated.
- **Else** the u'_k is deemed safe and applied to the plant.

If the model of the plant is known then $\mathcal{S}_k^+ := Ax_k \oplus Bu'_k \oplus \mathcal{W}$ (see Definition 6.1). However, based on Assumption 6.2, the system matrices A, B are unknown. Therefore, the forward one-step evolution of the system cannot be directly derived from (6.2). Thus, in what follows, by adopting the data-driven solution developed in [48, 88], we compute an outer approximation of $\mathcal{S}_k^+$, namely $\hat{\mathcal{S}}_k^+$, such that $\hat{\mathcal{S}}_k^+ \supseteq \mathcal{S}_k^+$. For the sake of completeness, the following two Lemmas summarize how $\hat{\mathcal{S}}_k^+$ can be computed in a data-driven fashion.

Lemma 6.1 *[48, Lemma 1] Let* $T = \sum_{i=1}^{N_t} N_s^{(i)}$ *and consider the following concatenation of multiple noise zonotopes*

$$\mathcal{M}_w = \mathcal{M}_w(C_w, [G_{M_w}^{(1)}, \dots, G_{M_w}^{(qT)}]),$$

where $C_w \in \mathbb{R}^{n\times(n+m)} = [c_w, \dots, c_w]$*, and* $G_{M_w} \in \mathbb{R}^{n\times T(n+m)}$ *is built* $\forall\, i \in \{1, \dots, q\}$, $\forall\, j \in \{2, \dots, T-1\}$ *as*

$$\begin{aligned} G_{M_w}^{(1+(i-1)T)} &= \begin{bmatrix} g_w^{(i)} & 0_{n\times(T-1)} \end{bmatrix} \\ G_{M_w}^{(j+(i-1)T)} &= \begin{bmatrix} 0_{n\times(j-1)} & g_w^{(i)} & 0_{n\times(T-j)} \end{bmatrix} \\ G_{M_w}^{(T+(i-1)T)} &= \begin{bmatrix} 0_{n\times(T-1)} & g_w^{(i)} \end{bmatrix} \end{aligned} \tag{6.14}$$

Then, the matrix zonotope

$$\begin{aligned}\mathcal{M}_{AB} &= (X_+ - \mathcal{M}_w)\begin{bmatrix} X_- \\ U_- \end{bmatrix}^\dagger \\ &:= \{[\hat{A},\, \hat{B}] : [\hat{A},\, \hat{B}] = C_{AB} + \sum_{i=1}^{T} \beta^{(i)} G^{(i)}_{M_{AB}}, \\ &\quad -1 \leq \beta^{(i)} \leq 1\}\end{aligned} \tag{6.15}$$

where $[\cdot]^\dagger$ *is the right pseudo inverse operator and*

$$\begin{aligned} C_{AB} &= (X_+ - C_w)\left([X_-^T,\, U_-^T]^T\right)^\dagger \\ G_{M_{AB}} &= \left[G^{(1)}_{M_w}\left([X_-^T,\, U_-^T]^T\right)^\dagger, \ldots, G^{(qT)}_{M_w}\left([X_-^T,\, U_-^T]^T\right)^\dagger\right] \end{aligned}$$

contains the set of all system matrices $[\hat{A},\, \hat{B}]$ *consistent with* (6.8) *and* $\mathcal{Z}_w$ *and such that* $[A, B] \in \mathcal{M}_{AB}$. □

Lemma 6.2 *(Adapted from [48, Theorem 1]) The set* $\hat{\mathcal{S}}_k^+ \subset \mathbb{R}^n$*, computed as*

$$\hat{\mathcal{S}}_k^+ = \mathcal{M}_{AB}[x'_k, u_k]^T \oplus \mathcal{W} \tag{6.16}$$

is a conservative outer approximation of $\hat{\mathcal{S}}_k^+$.

6.3.2 Emergency Controller Module

In the considered setup (see Section 6.2.5), the Emergency Controller does not have access to the reference signal r_k. Consequently, its objective is to preserve the plant's safety (during the attack) and to guarantee performance recovery (when the attack is terminated). By denoting the logic of the emergency controller as

$$u_k^e = f_e(x_k), \quad f_e : \mathcal{X}_e \subset \mathbb{R}^n \to \mathcal{U}_e \subset \mathbb{R}^m \tag{6.17}$$

(6.17) must be designed to fulfill the following requirements:

(1) Domain of attraction:

$$\mathcal{X} \supseteq \mathcal{X}_e \supseteq \mathcal{X}_\eta, \quad \mathcal{U}_e \subseteq \mathcal{U}, \tag{6.18}$$

(2) Finite-time Uniformly Ultimately Bounded (UUB) stability in a set

$$\hat{\mathcal{T}}_0 \subseteq \mathcal{X}_\eta \tag{6.19}$$

Condition (6.18) ensures that the emergency controller fulfills all the constraints and that it can be activated anytime and from any state reachable under the tracking controller, while condition (6.19) guarantees that, in the post-attack phase, the networked controller can be safely re-activated in a finite number of steps.

As shown in, e.g., [2], a controller satisfying the requirement above can be obtained by customizing (starting from an equilibrium point inside $\mathcal{X}_\eta$), the set-theoretic data-driven controller developed in [40]. Although effective, such a design completely ignores the current state vector of the plant. Consequently, the action of the emergency controller can take the state trajectory far from the desired reference signal. In what follows, we mitigate such a drawback by designing a data-driven set-theoretic emergency controller from a set of $L \geq 1$ admissible equilibrium points $(x_e^l, u_e^l), l \in \mathcal{L} := \{1, \cdots, L\}$. A pair (x_e^l, u_e^l) is considered admissible/safe if $x_e^l \in \mathcal{X}_\eta$, $\forall l$ and there exists a feedback controller

$$u_k^l = K_l(x_k - x_e^l) + u_e^l, \tag{6.20}$$

with gain $K_l \in \mathbb{R}^{m \times n}$, and such that the associated minimal RCI set (computed, e.g., using the data-driven methods developed in [91, 93] and [40, Remark 5]), namely $\hat{\mathcal{T}}_0^l \in \mathbb{R}^n$, is contained in the tracking controller's domain, i.e., $\hat{\mathcal{T}}_0^l \subseteq \mathcal{X}_\eta$.

Note that by construction, the L controllers (6.20) and associated RCI sets $\hat{\mathcal{T}}_0^l, l \in \mathcal{L}$ might not guarantee that the requirements (6.18),(6.19) are fulfilled, i.e., it might exist $x \in \mathcal{X}_\eta$ such that $x \notin \bigcup_{l=1}^L \hat{\mathcal{T}}_0^l = \mathcal{X}_e$. Therefore, to enlarge the domain $\mathcal{X}_e$ and comply with (6.18), (6.19) the following strategy is adopted.

- First, a Voronoi partition of $\mathcal{X}_\eta$ is created (see, e.g., the 5 partitions in Figure. 6.2). Therefore, a family of polyhedral regions $\{\mathcal{V}_l\}_{l=1}^L$ enjoying the following properties is obtained:

$$\mathcal{V}_l = \left\{ x \in \mathcal{X}_\eta : ||x - x_e^l||_2 \leq ||x - x_e^j||_2, \forall j \neq l, j \in \mathcal{L} \right\} \tag{6.21}$$

$$\bigcup_{l=1}^{L} \mathcal{V}_l = \mathcal{X}_\eta \tag{6.22}$$

- Then, by resorting to a data-driven set-theoretic predictive controller paradigm proposed in [40], we enlarge the Domain of Attraction (DoA) of each $l-th$ controller (6.20) to cover the associated Voronoi partition $\mathcal{V}_l$. To this end, a family of data-driven robustly controllable sets is recursively built by adapting the definition of ROSC sets (see Definition. 4.1) to the one-step evolution of model (6.3) under (6.4) in an offline phase. In particular, families $\left\{\hat{\mathcal{C}}_j^l\right\}_{j=1}^{N_l}$, of $N_l \geq 0$ of ROSC sets are built, with N_l satisfying the termination condition $\bigcup_{j=1}^{N_l} \left\{\hat{\mathcal{C}}_j^l\right\} = \mathcal{V}_l, \forall l = 1, \cdots, L$.

As long as the data-driven computation of $\{\hat{\mathcal{C}}_j^l\}_{j=1}^{N}$, is concerned, we resort to the procedure, summarized in Lemma 6.3, which resorts to an augmented description of the ROSC sets.

Lemma 6.3 *Consider a collection of input-state trajectories for* (6.3)-(6.4) *fulfilling Assumption 6.2. an inner approximation of the ROSC set 4.1 can be computed as follows [40, Sec. III.C]:*

$$\begin{aligned} \hat{\mathcal{C}}_j^l &= Proj_x(\hat{\Xi}_j^l) = \left\{x \in \mathbb{R}^n : H_{\hat{\mathcal{C}}_j^l} x \leq h_{\hat{\mathcal{C}}_j^l}\right\}, \\ \hat{\Xi}_j^l &= In_z\left\{\hat{\Xi}_{AB}^l\right\}, \\ \hat{\Xi}_{j_{AB}}^l &= \bigcap_{[\hat{A}_i, \hat{B}_i] \in \mathcal{V}_{AB}} \left\{z = [x^T, u^T]^T \in \mathbb{R}^{n+m} : H_i^l z \leq h_i^l\right\} \end{aligned} \tag{6.23}$$

where $\hat{\Xi}_{j_{AB}}^l$ is the $(x,u)-$ augmented description of the ROSC set, $\mathcal{V}_{AB}$ denotes the matrix vertices of $\mathcal{M}_{AB}$, $In_z(\cdot)$ is an operator computing a zonotopic inner approximation of a polytope, $Proj_x(\hat{\Xi}_j^l)$ performs a projection of $\hat{\Xi}_j^l$ into the $x-$domain, and

$$H_i^l = \begin{bmatrix} H_x & 0 \\ H_{\hat{\mathcal{C}}_{j-1}^l} \hat{A}_i & H_{\hat{\mathcal{C}}_{j-1}^l} \hat{B}_i \\ 0 & H_u \end{bmatrix}, \quad h_z^i = \begin{bmatrix} h_x \\ \tilde{h}_{\hat{\mathcal{C}}_{j-1}^l} \\ h_u \end{bmatrix} \tag{6.24}$$

with

$$[\tilde{h}_{\hat{\mathcal{C}}_{j-1}^l}]_r = \min_{w \in \mathcal{W}} \left\{[h_{\hat{\mathcal{C}}_{j-1}^l}]_r - [H_{\hat{\mathcal{C}}_{j-1}^l}]_r w\right\} \tag{6.25}$$

and $[h_{\hat{\mathcal{C}}^l_{j-1}}]_r$, $[H_{\hat{\mathcal{C}}^l_{j-1}}]_r$ *the* $r-th$ *rows of* $h_{\hat{\mathcal{C}}^l_{j-1}}$ *and* $H_{\hat{\mathcal{C}}^l_{j-1}}$.

□

Given $\{\hat{\mathcal{C}}^l_j\}_{j=0}^N$, and a convex cost function $J(x_k, u)$, the online operations of the Safety Verification Module and Emergency data-driven set-theoretic controller can be summarized as in Algorithm 6.1. where the flag f is used to make sure that the emergency controller,

Algorithm 6.1 Safety Verification (SV) and Emergency Controller (EC)

Offline: Compute $\hat{\mathcal{C}}^l_0$ and $\{\hat{\Xi}^l_j, \hat{\mathcal{C}}^l_j\}_{j=1}^{N_l}, l{=}1, \ldots, L$ as in (6.23). Set $f = 1$.

Online ($\forall\, k$)*:*

if $f == 0$ **or** $u'_k \notin \mathcal{U}$ **or** $\mathcal{S}_k^+ \not\subseteq \mathcal{X}_\eta$ **then** ▷ *Activate EC*

▷ *EC starts*

  Set $f = 0$ and find $\bar{l} \in \mathcal{L}$ such that $x_k \in \mathcal{V}_{\bar{l}}$

  Find $\bar{j}_k := \min_{j \in \{0,\ldots,N_l\}} \{j : x_k \in \hat{\mathcal{C}}^l_j\}$

  if $\bar{j}_k == 0$ **then**

    $u^e_k = f^l_0(x_k)$, $f = 1$

  else

$$u^e_k = \arg\min_u J(x_k, u) \quad s.t. \left[x_k^T, u^T\right]^T \in \hat{\Xi}^l_{j_k} \tag{6.26}$$

  end if ▷ *EC ends*

  Apply u^e_k ▷ *EC control law*

else

  Apply u'_k ▷ *TS control law*

end if

once activated, will remain active at least until the terminal region is reached. Given the operations of the SV and EC modules (see Algorithm 6.1), the following proposition holds true (adapted from the results in [35, Proposition 1]):

Proposition 6.1 *[35] Consider the sets of equilibrium pairs* $\left\{(x^l_e, u^l_e)\right\}$ *and Voronoi partition* $\{\mathcal{V}_l\}_{l=1}^L$*, the RCI sets* $\left\{\mathcal{T}^l_0\right\}_{l=1}^L$*, and the families of ROSC sets* $\left\{\hat{\mathcal{C}}^l_j\right\}_{j=1}^{N_l}, l \in \mathcal{L}$ *computed according to* (6.23)*. Then, if at* $k = k'$ *a persistent cyber-attack starts on the actuation channel and* $x_{k'} \in \mathcal{V}_l, 1 \leq l \leq L$*, then the emergency controller ensures that safety and recovery are guaranteed (see Definition 6.2). Moreover, for* $k \geq k' + N_l$ *the tracking error* $e_k = x_k - r_k$ *is such that* $e_k \leq d^{sup}(\hat{\mathcal{T}}^{l_i}_0, r_k), \forall k \geq k' + N_l$.

6.3.3 Tracking Supervisor Module (*TS*)

If the anomaly detector module detects an anomaly, then the attack could be either on the actuation and/or measurement channel. In what follows, the *TS* actions are derived assuming that the cyber-attack affects the measurement channel. In particular, the objective of this module is to allow the networked control system to operate safely in an open-loop mode and minimize the tracking performance loss.

By taking into account the worst-case attack detection delay τ, if a cyber-attack is detected at k', then the last reliable measurement is $x_{k'-\tau-1}$. However, by exploiting forward reachability arguments, it is possible to robustly estimate from $x_{k'-\tau-1}$ the set of states $\hat{\mathcal{R}}_{k'}$ containing the current state $x_{k'}$, i.e., such that $x_{k'} \in \hat{\mathcal{R}}_{k'}$. In a data-driven fashion, the set $\hat{\mathcal{R}}_{k'}$ can be outer approximated by resorting to the following recursive RORS data-driven computation starting from the initial condition $\hat{\mathcal{R}}_{k'-\tau-1} = x_{k'-\tau-1}$,

$$\hat{\mathcal{R}}_{k'-\tau+t} = \mathcal{M}_{AB}\left[\mathcal{R}_{k'-\tau+t-1}, u_{k'-\tau+t-1}\right] \oplus \mathcal{W}, \quad \forall\, t \in \mathbb{Z}_+ \tag{6.27}$$

Given $\hat{\mathcal{R}}_{k'-\tau+t}$, *TS* is instructed to replace $\forall k \geq k'$, x_k with an admissible state $\hat{x}_k \in \hat{\mathcal{R}}_k$. Consequently, under attack, the tracking controller will compute the following control action

$$u_k = \eta(\hat{x}_k, r_k), \quad \forall\, k \geq k' \tag{6.28}$$

6.3.4 Tracking Performance Evaluation

To design the tracking supervisor logic to minimize the tracking performance loss under attack, we need to first offline approximately quantify the tracking performance degradation associated with the emergency controller actions. In particular, the tracking index $I(i,j)$ is here proposed:

$$I(i,j) = \alpha I_1(l_i, l_j) + \beta I_2(l_i, l_j), \quad l_i, l_j \in \mathcal{L} \tag{6.29}$$

where $\alpha, \beta \geq 0$ are two weighting factors and

- $I_1(l_i, l_j) = d^{sup}(\hat{\mathcal{T}}_0^{l_i}, x_e^{l_j})$ and $d^{sup}(\mathcal{S}, p)$ computes the maximum distance between a point $p \in \mathbb{R}^s$ and set $\mathcal{S} \subset \mathbb{R}^s$ (see [35, Definition 2]). Such an index quantifies the

nominal tracking error if $x \in \hat{\mathcal{T}}_0^{l_i} \subseteq \mathcal{V}_{l_i}$ and $r_{k'}$ belongs to $\mathcal{V}_{l_j}$. In particular, $\hat{\mathcal{T}}_0^{l_i}$ is the RCI set where the state of the system will be confined in N_{l_i} steps if the emergency controller is activated at the current time, and $x_e^{l_j}$ is the disturbance-free equilibrium point of the partition containing r_k.

- $I_2(l_i, l_j) = \min_{0 \leq p \leq N_{l_j}} p : \hat{\mathcal{T}}_0^{l_i} \subseteq \bigcup_{s=0}^{p} \{\hat{\mathcal{C}}_s^{l_j}\}$, with $\{\hat{\mathcal{C}}_s^{l_j}\}_{s=0}^{N_{l_j}}$ a set of $N_{l_j} \geq 0$ ROSC set built as prescribed by (4.4) with starting RCI set $\hat{\mathcal{T}}_0^{l_j} = \mathcal{V}_j$ and terminal condition $\hat{\mathcal{T}}_0^{l_i} \subseteq \bigcup_{s=0}^{N_{l_j}} \{\hat{\mathcal{C}}_s^{l_j}\}$. Such index quantifies the worst-case number of steps required for $x_k \in \hat{\mathcal{T}}_0^{l_i} \subseteq \mathcal{V}_{l_i}$ to enter the Voronoi partition $\mathcal{V}_{l_j}$ containing $r_{k'}$.

Remark 6.2 *In simpler terms, by assuming a constant reference signal during the attack phase, $I_1(l_i, l_j)$ approximates the steady-state tracking error committed activating the emergency controller during the attack, while $I_2(l_i, l_j)$ approximates the time required to recover the reference tracking problem when the attack is terminated.* □

Then, it is possible to sort all the pairs $(l_i, l_r), \forall l_i \in \mathcal{L}$ in an ascending order according to the tracking index $I(l_i, l_r)$, i.e.,

$$\begin{aligned} \mathcal{I}(l_r) &= \Big[I(l_1, l_r), \cdots, I(l_x, l_r), \cdots, I(l_L, l_r) \Big] \\ l_j &\in \mathcal{L}, \forall j, I(l_1, l_r) \leq \cdots \leq I(l_x, l_r), \leq \cdots, \leq I(l_L, l_r) \end{aligned} \tag{6.30}$$

Therefore, if $I(l_x, l_r) = I(l_1, l_r)$ then, the lowest tracking performance loss is obtained by forcing x'_k, $\forall\, k \geq k'$, to remain in $\mathcal{V}_{l_x}$. On the other hand, if $I(l_x, l_r) \neq I(l_1, l_r)$ then better tracking performance are obtained if $x'_k, k \geq k'$ can be steered into a pair $I(l_j, l_r)$ such that $I(l_j, l_r) < I(l_x, l_r)$.

Since we can only robustly predict the robust reachable sets, $\hat{\mathcal{R}}_k$ prevents us from deterministically evaluating the tracking index $I(l_i, l_j)$ at the next time instant using a single vector approach. Consequently, the following index J is defined:

$$J_{k+1} = \sum_{l_j \in \mathcal{L}} \frac{vol\left(\hat{\mathcal{R}}_{k+1} \cap \mathcal{V}_{l_j}\right)}{vol\left(\hat{\mathcal{R}}_{k+1}\right)} I(l_j, l_{r_k}), \quad k \geq k' \tag{6.31}$$

where $vol(\cdot)$ computes the volume of a set and J_{k+1} defines a weighted sum of the tracking

index based on the volume overlap between the uncertain prediction set and the Voronoi regions. Then, the open-loop tracking controller is kept active until one of the following stopping conditions is verified:

(1) $\hat{\mathcal{R}}_{k+1} \notin \mathcal{X}_\eta$. Such a condition implies that u_k, if applied to the plant, could bring the state of the plant outside of the admissible controller regions

(2) $J_{k+1} > J_k$. This condition implies that the robust tube containing the state trajectory is moving toward regions with higher tracking performance loss (according to the index I).

When one of the two conditions above arises, the *TS* is instructed to replace u_k with an invalid $u_k \in \mathcal{U}$ to activate the emergency controller intentionally (see Section 6.3.1). Given the above results, the logic of the tracking supervisor has been summarized in Algorithm. 6.2.

Algorithm 6.2 Data-Driven Tracking Supervisor (*TS*)

Online ($\forall\, k$)*:*

if $d_k == 1$ **then**
  Estimate $\hat{\mathcal{R}}_k$ using (6.27) and compute u_k as in (6.28).
  if ($\hat{\mathcal{R}}_{k+1} \not\subseteq \mathcal{X}_\eta$) or ($J_{k+1} > J_k$) **then**
    Replace u_k with any $u_k \notin \mathcal{U}$.
  end if
end if
u_k is sent.

Proposition 6.2 *Under any admissible cyber-attack, the data-driven tracking supervisor (Algorithm 6.2) allows obtaining a tracking performance index I better or equal to the one obtainable using only the safety verification module and emergency controller (Algorithm 6.1).*

Proof 6.1 *In the worst-case scenario, a cyber-attack might not be detected by the used anomaly detector module* (6.7) *or the attack corrupt the actuation channel. In these two cases, the TS is never activated, or its actions are invalidated by the attack on the actuation channel. Consequently, the tracking performance index I depends only on the actions of the Algorithm 6.1. On the other hand, if the cyber-attack affects the measurement channel, then*

Algorithm 6.2 can allow the plant to obtain a better performance index I. Assume that the attack starts at k' and that one of the TS stopping conditions is valid at $k' + k_{stop}$, with $k_{stop} \in \mathbb{Z}_+$. Then, TS will allow an open-loop tracking of r_k for $k' \leq k < k' + k_{stop}$. As a consequence, since a stop condition arises only at $k' + k_{stop}$, it is also true that $J_{k+1} \leq J_k, \forall\, k' \leq k < k' + k_{stop}$. Therefore, the latter implies that the tube containing the state trajectory of the plant is moving towards a Voronoi partition with an associated tracking index better than the current one. In the worst-case scenario, if $k_{stop} = 0$ (i.e., the open-loop controller cannot be used at all), then the TS module is instructed to replace u_k with an invalid $u_k \notin \mathcal{U}$, so causing the activation of the EC. ■

Remark 6.3 *The proposed solution has been developed assuming a detection delay $\tau > 0$. However, in some application scenarios, the communication channels between the plant and the controller are authenticated, i.e., a Message Authentication Code (MAC) [80] is used to authenticate every data packet sent over the network. Such a security mechanism allows the controller (or plant) to verify the authenticity and integrity of the received sensor measurements (or control actions), hence allowing instantaneous detection of network attacks. In this setup, the proposed solution can be straightforwardly adapted by simply setting $\tau = 0$.*

6.4 Simulation Results

In this section, we consider the industrial Continuous-Stirred Tank Reactor (CSTR) system used in [97] as a testbed to evaluate the proposed approach. In this plant, chemical species S_A reacts to form species S_B, the state vector and control inputs are $x_p = [C_A, T]^T$ and $u = [T_C, C_{Ai}]^T$, where C_A is the concentration of S_A in the tank, T the reaction temperature, T_C the cooling medium temperature, and C_{Ai} is input concentration of S_A. A linearized discrete-time representation of the CSTR system using a sampling time $T_s = 1$ sec has been provided in [97], and it is characterized by the system matrices

$$A = \begin{bmatrix} 0.9719 & 0.0013 \\ 0.0340 & 0.8628 \end{bmatrix}, B = \begin{bmatrix} -0.0839 & 0.0232 \\ 0.0761 & 0.4144 \end{bmatrix} \tag{6.32}$$

which we assume to be apriori unknown according to the considered data-driven setup. By considering a bounded disturbance set $\mathcal{W} = \{w : [-0.001, -0.001]^T \leq w \leq [0.001, 0.001]^T\}$ and assuming that the state and input constraints are $-2 \leq T_C \leq 2, -10 \leq C_{Ai} \leq 10, -10 \leq C_A \leq 10, -30 \leq T \leq 30$, we have simulated the CSTR system for random input perturbation. Consequently, we have collected 4 input-state trajectories that verify the rank condition (6.11). The estimated data-driven model $\mathcal{M}_{AB}$ is available on the provided GitHub link (see Section 6.1.1).

The emergency controller is configured with a five region Voronoi partition of $\mathcal{X}_\eta$ obtained using as generators the equilibrium states $x_e^1 = [4, 15]^T, x_e^2 = [-6, 15]^T, x_e^3 = [0, 0]^T, x_e^4 = [6, -20]^T, x_e^5 = [-4, -20]^T$. On the other hand, the tracking supervisor is configured to use $\alpha = 1, \beta = 0$ in (6.29). Moreover, we have used the proposed data-driven anomaly detector introduced in [2] and by simulating the system under different false-data injection attacks, a worst-case detection delay of $\tau = 5$ has been obtained. In the performed simulations, $x_0 = [0.01, -0.01]^T$ and the plant is required to track the time-varying reference signal r_k shown in Figure 6.2 (see the blue stars), while three different cyber-attacks on the measurement channel occur. The first attack, for $60 \leq k \leq 110$, injects $x_k^a = 0.01[k - 59, k - 59]^T$ on the measurement vector x_k. The anomaly detector identifies the presence of an attack with a delay at $k = 62$ and activates the tracking supervisor ($d_k = 1$). Right after the attack has been detected, the tracking supervisor is instructed to assume x'_{56} as the last valid measurement to compute the predicted robust forward reachable sets (see red regions in Figure. 6.2). The tracking supervisor actions proceed until $k = 106$, where $\hat{\mathcal{R}}_{107} \not\subseteq \mathcal{X}_\eta$ and for safety reasons, the control input is invalidated, and the emergency controller is activated. As a consequence for $107 \leq k < 129$, the emergency controller steers the state of the system into the RCI region $\mathcal{T}_0^1$ centered in x_e^1, and the reference tracking task is temporarily paused. However, when the first attack ends (at $k = 110$), the plant recovers its tracking task (see Figures. 6.2-6.3). In the second scenario, an attack intermittently affects the measurement channel for $200 \leq k < 261$. In this case, the attacker injects $x_k^a = 0.08[k - 199, k - 199]^T, x_k^a = 0.1[k - 239, k - 239]^T$ during $k \in [200, 220]$ and $k \in [240, 260]$, respectively. The first attack is detected at $k = 202$, and the second attack is detected at $k = 241$. However, due to the nature of the attack, the sporadically

received measurements enable the tracking supervisor to reset the uncertainty set (yellow sets in Figure. 6.2), thereby preventing the suspension of the tracking task. In the third attack scenario, the attacker injects $x_k^a = 0.1[k-399,\, k-399]^T$ on the measurements for $400 \leq k \leq 420$. In this case, the anomaly detector identifies the presence of an attack at $k = 401$ and activates the tracking supervisor accordingly. Differently from the first attack scenario, in this case, the uncertain predicted forward one-step evolution sets estimated by the tracking supervisor never violate the safety constraints (green sets in Figure. 6.3). Moreover, the index J presents a decreasing behavior, denoting that the tracking performance is improving under the action of the open-loop tracking controller's actions (see green sets in Figure. 6.2). As a consequence, the tracking task is never suspended under the actions of the cyber-attack. Finally, in Figure. 6.3 and Table. 6.1, the proposed solution is contrasted with the one in [2]. Since in [2], the emergency controller is activated regardless of the nature of the attack, an unavoidable tracking loss occurs in all the three considered attacks with a consequence of bigger tracking performance loss. By measuring the tracking error, namely e_r, as $e_r = \sum_{k=1}^{N_s} \frac{||x_k - r_k||}{N_s}$, with N_s the simulation steps, Table 6.1 reports the obtained numerical results. It is possible to appreciate how, compared to [2], the proposed solution reduces the tracking performance degradation due to the presence of cyber-attacks.

Table 6.1: Tracking Error: proposed approach, [2], no attack

	No attack	Proposed Approach	[2]
e_r	1.2186	1.5878	6.0533

6.5 Conclusions

This chapter proposed a data-driven robust solution to the safety and reference tracking control problems for constrained CPSs by leveraging robust reachability arguments. The proposed control architecture included two data-driven add-on modules (local to the plant and to the networked controller) designed to ensure safety while enhancing, whenever possible, the tracking performance under cyber-attacks in a supervised manner. Theoretical and simulation results have been reported to show the effectiveness of the proposed scheme.

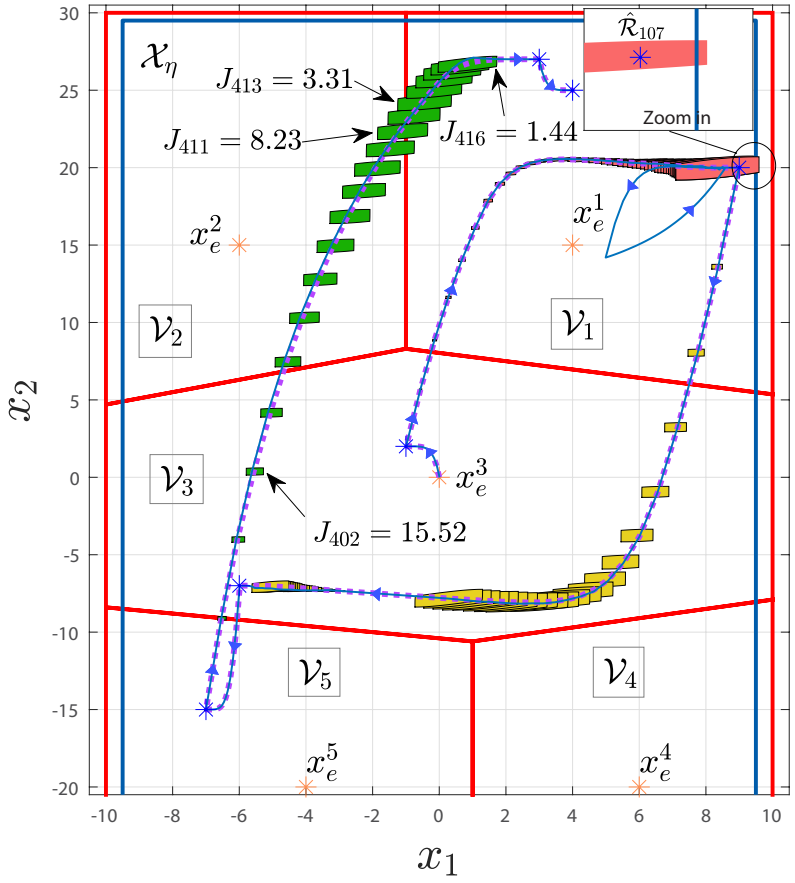


Figure 6.2: State trajectory: proposed solution with attacks (blue solid line) vs trajectory in attack-free scenario (purple dashed line).

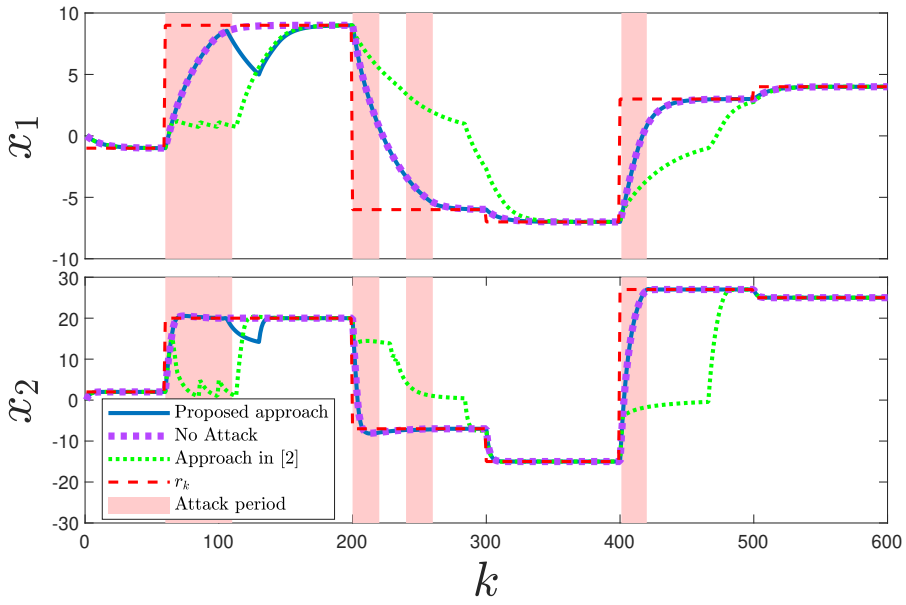


Figure 6.3: State evolution: no attack, proposed approach, [2].

Chapter 7

Conclusion and Future Works

In this thesis, data-driven methods for the safety and security of constrained CPSs have been investigated. For this purpose, we first discussed a novel active detection mechanism based on the concept of dimensionality reduction and PCA to detect intelligent and coordinated cyber attacks in networked control systems. Then, we presented a data-driven approach to compute backward reachable sets and ST-MPC. Afterward, we presented data-driven control architectures for preserving the safety and tracking performance of constrained CPSs against cyber attacks.

In chapter 3, we designed an encoding mechanism to encode the sensor outputs into a randomly changing lower-dimensional space obtained using principal component analysis. Such a transformation ensures optimal information reconstruction on the controller's side, and it prevents the attacker from accessing the original sensor measurements, nullifying the possibility of perfect stealthy attacks. Then, on the controller's side, we design a passive Gaussian anomaly detector that leverages the output of an unknown input observer ad-hoc designed to estimate the system's states and unknown inputs simultaneously. It is formally shown that the proposed detection strategy is able to discover the presence of replay and covert attacks.

In chapter 4, we proposed a new method for computing robust backward reachable sets from noisy data for unknown constrained linear systems affected by bounded disturbances. Our approach introduces an algorithm that generates zonotopic inner approximations, which can be applied in control design. It is demonstrated that these sets, when constructed in an

extended space that includes both states and inputs, can represent the system's one-step dynamics within the computed extended region. This result is leveraged to create a set-theoretic model predictive controller that, in an offline phase, constructs a recursive family of robust data-driven reachable sets, and in real-time, determines admissible control actions recursively without explicitly relying on a system model or the available data.

In chapter 5, we introduced a data-driven networked control framework for unknown and constrained cyber-physical systems that can detect false-data-injection attacks and preserve the plant's safety. Specifically, on the controller side, we develop a novel robust anomaly detector that identifies network attacks by utilizing a data-driven outer approximation of the expected robust one-step reachable set. Meanwhile, on the plant side, we create a data-driven safety verification module that applies worst-case analysis to assess whether the received control input is safe for the plant's operation. If needed, this module also switches to a local data-driven set-theoretic model predictive controller, ensuring the plant remains within a predefined safe state until the attack is resolved.

In chapter 6, a worst-case data-driven control framework proposed to ensure the safety of constrained cyber-physical systems under cyber-attacks while minimizing performance degradation in tracking whenever possible. To this end, a data-driven robust anomaly detector has been developed to identify cyber-attacks. Additionally, a tracking supervisor module has introduced to enable safe open-loop tracking control when measurements become unreliable. On the plant side, a safety verification module and a local emergency controller have been designed to handle severe attack scenarios that cannot be addressed by the controller. These modules use worst-case reachability and controllability data-driven strategies to detect unsafe conditions and, when necessary, replace the tracking controller with emergency actions aimed at guiding the plant's state trajectory into a predefined safe and admissible control-invariant region until the attack is resolved.

7.1 Future research directions

Potential future research directions to extend and enhance the work presented in this thesis are outlined below:

- The active detection mechanism developed in chapter 3, can be enhanced with data-driven attack mitigation strategies that are activated once an attack is revealed. As a result, the safety and tracking performance of the plant can be preserved.

- The proposed data-driven approach in chapter 4 can be improved by mitigating the computational complexity and conservativeness of the proposed procedure to compute ROSC sets.

- The safety and tracking performance preserving architectures proposed Chapter 5 and 6 can be improved to reduce their conservativeness. Moreover, the used technique can be extended to deal with complex plant models with, e.g., nonlinear or piece-wise affine dynamics.

- The theoretical results presented in this thesis can be validated through implementation on real-world CPSs.

Bibliography

[1] F. Miao, Q. Zhu, M. Pajic, and G. J. Pappas, "Coding schemes for securing cyber-physical systems against stealthy data injection attacks," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 1, pp. 106–117, 2016.

[2] M. Attar and W. Lucia, "A data-driven safety preserving control architecture for constrained cyber-physical systems," *International Journal of Robust and Nonlinear Control*, 2024, doi: 10.1002/rnc.7654.

[3] D. Mortari, "n-dimensional cross product and its application to the matrix eigenanalysis," *Journal of Guidance, Control, and Dynamics*, vol. 20, no. 3, pp. 509–515, 1997.

[4] Y. Mo, T. H.-J. Kim, K. Brancik, D. Dickinson, H. Lee, A. Perrig, and B. Sinopoli, "Cyber–physical security of a smart grid infrastructure," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 195–209, 2011.

[5] S. Amin, X. Litrico, S. S. Sastry, and A. M. Bayen, "Stealthy deception attacks on water scada systems," in *Proceedings of the 13th ACM international conference on Hybrid systems: computation and control*, 2010, pp. 161–170.

[6] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE transactions on automatic control*, vol. 58, no. 11, pp. 2715–2729, 2013.

[7] E. Bou-Harb, W. Lucia, N. Forti, S. Weerakkody, N. Ghani, and B. Sinopoli, "Cyber meets control: A novel federated approach for resilient cps leveraging real cyber threat intelligence," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 198–204, 2017.

[8] A. Barboni, H. Rezaee, F. Boem, and T. Parisini, "Detection of covert cyber-attacks in interconnected systems: A distributed model-based approach," *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3728–3741, 2020.

[9] D. Ding, Q.-L. Han, Y. Xiang, X. Ge, and X.-M. Zhang, "A survey on security control and attack detection for industrial cyber-physical systems," *Neurocomputing*, vol. 275, pp. 1674–1683, 2018.

[10] L. Niu, Z. Li, and A. Clark, "Lqg reference tracking with safety and reachability guarantees under false data injection attacks," in *2019 American Control Conference (ACC)*. IEEE, 2019, pp. 2950–2957.

[11] V. S. Dolk, P. Tesi, C. De Persis, and W. Heemels, "Event-triggered control systems under denial-of-service attacks," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 1, pp. 93–105, 2016.

[12] A. A. Cardenas, S. Amin, and S. Sastry, "Secure control: Towards survivable cyber-physical systems," in *International Conference on Distributed Computing Systems Workshops*. IEEE, 2008, pp. 495–500.

[13] M. Ghaderi, K. Gheitasi, and W. Lucia, "A blended active detection strategy for false data injection attacks in cyber-physical systems," *IEEE Transaction on Control of Network Systems*, vol. 8, no. 1, pp. 168–176, 2020.

[14] J. Hespanha, P. Naghshtabrizi, and Y. Xu, "A survey of recent results in networked control systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 138–162, 2007.

[15] T. Chen, "Stuxnet, the real start of cyber warfare?[editor's note]," *IEEE Network*, vol. 24, no. 6, pp. 2–3, 2010.

[16] N. Kshetri and J. Voas, "Hacking power grids: a current problem," *Computer*, vol. 50, no. 12, pp. 91–95, 2017.

[17] A. Teixeira, D. Pérez, H. Sandberg, and K. Johansson, "Attack models and scenarios for networked control systems," in *Proceedings of the 1st international conference on High Confidence Networked Systems*. ACM, 2012, pp. 55–64.

[18] 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.

[19] G. Carl, G. Kesidis, R. R. Brooks, and S. Rai, "Denial-of-service attack-detection techniques," *IEEE Internet computing*, vol. 10, no. 1, pp. 82–89, 2006.

[20] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *Allerton conference on communication, control, and computing*. IEEE, 2009, pp. 911–918.

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

[22] Y. Shoukry, P. Martin, P. Tabuada, and M. Srivastava, "Non-invasive spoofing attacks for anti-lock braking systems," in *Cryptographic Hardware and Embedded Systems-CHES 2013: 15th International Workshop, Santa Barbara, CA, USA, August 20-23, 2013. Proceedings 15*. Springer, 2013, pp. 55–72.

[23] S. Sridhar, A. Hahn, and M. Govindarasu, "Cyber–physical system security for the electric power grid," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 210–224, 2011.

[24] C. Schellenberger and P. Zhang, "Detection of covert attacks on cyber-physical systems by extending the system dynamics with an auxiliary system," in *IEEE Conference on Decision and Control (CDC)*. IEEE, 2017, pp. 1374–1379.

[25] S. Weerakkody and B. Sinopoli, "Detecting integrity attacks on control systems using a moving target approach," in *IEEE Conference on Decision and Control (CDC)*. IEEE, 2015, pp. 5820–5826.

[26] W. Lucia, G. Franzè, and B. Sinopoli, "A supervisor-based control architecture for constrained cyber-physical systems subject to network attacks," *IEEE Transaction on Control of Network Systems*, vol. 10, no. 3, pp. 1184–1194, 2023.

[27] K. Gheitasi and W. Lucia, "A safety preserving control architecture for cyber-physical systems," *International Journal of Robust and Nonlinear Control*, vol. 31, no. 8, pp. 3036–3053, 2021.

[28] M. Pajic, I. Lee, and G. Pappas, "Attack-resilient state estimation for noisy dynamical systems," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 1, pp. 82–92, 2017.

[29] L. An and G. Yang, "Secure state estimation against sparse sensor attacks with adaptive switching mechanism," *IEEE Trans. on Automatic Control*, 2017.

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

[31] M. Zhu and S. Martinez, "On the performance analysis of resilient networked control systems under replay attacks," *IEEE Transactions on Automatic Control*, vol. 59, no. 3, pp. 804–808, 2014.

[32] L. An and G.-H. Yang, "Improved adaptive resilient control against sensor and actuator attacks," *Information Sciences*, vol. 423, pp. 145–156, 2018.

[33] Y. Yuan, H. Yuan, L. Guo, H. Yang, and S. Sun, "Resilient control of networked control system under dos attacks: A unified game approach," *IEEE transactions on Industrial Informatics*, vol. 12, no. 5, pp. 1786–1794, 2016.

[34] M. Zhu and S. Martinez, "On the performance analysis of resilient networked control systems under replay attacks," *IEEE Transactions on Automatic Control*, vol. 59, no. 3, pp. 804–808, 2013.

[35] K. Gheitasi and W. Lucia, "A worst-case approach to safety and reference tracking for cyber-physical systems under network attacks," *IEEE Transaction on Automatic Control*, vol. 68, no. 7, pp. 4391–4397, 2023.

[36] W. Liu, J. Sun, G. Wang, F. Bullo, and J. Chen, "Data-driven resilient predictive control under denial-of-service," *IEEE Transactions on Automatic Control*, vol. 68, no. 8, pp. 4722–4737, 2022.

[37] K. P. Wabersich and M. N. Zeilinger, "Linear model predictive safety certification for learning-based control," *IEEE Conference on Decision and Control (CDC)*, pp. 7130–7135, 2018.

[38] M. Bajelani, W. Lucia, and K. van Heusden, "A modular safety filter for safety-certified cyber-physical systems," *arXiv preprint arXiv:2403.15854*, 2024.

[39] M. Attar and W. Lucia, "An active detection strategy based on dimensionality reduction for false data injection attacks in cyber-physical systems," *IEEE Transaction on Control of Network Systems*, vol. 10, no. 4, pp. 1844–1854, 2023.

[40] ——, "Data-driven robust backward reachable sets for set-theoretic model predictive control," *IEEE Control Systems Letters*, vol. 7, pp. 2305–2310, 2023.

[41] Y.-C. Sun and G.-H. Yang, "Robust event-triggered model predictive control for cyber-physical systems under denial-of-service attacks," *International Journal of Robust and Nonlinear Control*, vol. 29, no. 14, pp. 4797–4811, 2019.

[42] Z. Zhao and Y. Xu, "Performance based attack detection and security analysis for cyber-physical systems," *International Journal of Robust and Nonlinear Control*, vol. 33, no. 5, pp. 3267–3284, 2023.

[43] Z.-S. Hou and Z. Wang, "From model-based control to data-driven control: Survey, classification and perspective," *Information Sciences*, vol. 235, pp. 3–35, 2013.

[44] C. De Persis and P. Tesi, "Formulas for data-driven control: Stabilization, optimality, and robustness," *IEEE Transaction on Automatic Control*, vol. 65, no. 3, pp. 909–924, 2019.

[45] J. Bongard, J. Berberich, J. Köhler, and F. Allgöwer, "Robust stability analysis of a simple data-driven model predictive control approach," *IEEE Transaction on Automatic Control*, vol. 68, no. 5, pp. 2625–2637, 2023.

[46] J. Berberich, J. Köhler, M. A. Müller, and F. Allgöwer, "Data-driven model predictive control with stability and robustness guarantees," *IEEE Transaction on Automatic Control*, vol. 66, no. 4, pp. 1702–1717, 2021.

[47] K. P. Wabersich and M. N. Zeilinger, "A predictive safety filter for learning-based control of constrained nonlinear dynamical systems," *Automatica*, vol. 129, p. 109597, 2021.

[48] A. Alanwar, A. Koch, F. Allgöwer, and K. H. Johansson, "Data-driven reachability analysis from noisy data," *IEEE Transactions on Automatic Control*, vol. 68, no. 5, pp. 3054–3069, 2023.

[49] M. Attar and W. Lucia, "A data-driven approach to preserve safety and reference tracking for constrained cyber-physical systems under network attacks," 2024. [Online]. Available: https://arxiv.org/abs/2410.00208

[50] G. Dan and H. Sandberg, "Stealth attacks and protection schemes for state estimators in power systems," *IEEE SmartGridComm*, pp. 214–219, 2010.

[51] Y. Mo and B. Sinopoli, "Integrity attacks on cyber-physical systems," in *Int. Conf. on High Confidence Networked Systems*. ACM, 2012, pp. 47–54.

[52] M. Bishop, "The art and science of computer security," 2002.

[53] S. Ding, *Model-based fault diagnosis techniques: design schemes, algorithms, and tools*. Springer Science & Business Media, 2008.

[54] I. Hwang, S. Kim, Y. Kim, and C. Seah, "A survey of fault detection, isolation, and reconfiguration methods," *IEEE transactions on control systems technology*, vol. 18, no. 3, pp. 636–653, 2010.

[55] Y. Chen, S. Kar, and J. Moura, "Dynamic attack detection in cyber-physical systems with side initial state information," *IEEE Transactions on Automatic Control*, vol. 62, no. 9, pp. 4618–4624, 2017.

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

[57] D. Angeli, A. Casavola, G. Franzè, and E. Mosca, "An ellipsoidal off-line mpc scheme for uncertain polytopic discrete-time systems," *Automatica*, vol. 44, no. 12, pp. 3113–3119, 2008.

[58] F. Borrelli, A. Bemporad, and M. Morari, *Predictive control for linear and hybrid systems*. Cambridge University Press, 2017.

[59] S. V. Rakovic, E. C. Kerrigan, D. Q. Mayne, and J. Lygeros, "Reachability analysis of discrete-time systems with disturbances," *IEEE Trans. on Automatic Control*, vol. 51, no. 4, pp. 546–561, 2006.

[60] M. Herceg, M. Kvasnica, C. N. Jones, and M. Morari, "Multi-parametric toolbox 3.0," in *2013 European control conference (ECC)*. IEEE, 2013, pp. 502–510.

[61] M. Althoff, "Reachability analysis and its application to the safety assessment of autonomous cars," *Ph.D. dissertation, Technische Universität München, München, Germany*, 2010.

[62] G. Basile and G. Marro, "On the observability of linear, time-invariant systems with unknown inputs," *Journal of Optimization theory and applications*, vol. 3, no. 6, pp. 410–415, 1969.

[63] J. Chen and R. J. Patton, *Robust model-based fault diagnosis for dynamic systems*. Springer Science & Business Media, 2012, vol. 3.

[64] S. Hui and S. Żak, "Observer design for systems with unknown inputs," *International Journal of Applied Mathematics and Computer Science*, vol. 15, pp. 431–446, 2005.

[65] G. Fiore, Y. H. Chang, Q. Hu, M. D. Di Benedetto, and C. J. Tomlin, "Secure state estimation for cyber physical systems with sparse malicious packet drops," in *American Control Conference (ACC)*. IEEE, 2017, pp. 1898–1903.

[66] Q. Hu, D. Fooladivanda, Y. H. Chang, and C. J. Tomlin, "Secure state estimation and control for cyber security of the nonlinear power systems," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 1310–1321, 2017.

[67] T. Yang, C. Murguia, M. Kuijper, and D. Nesic, "An unknown input multi-observer approach for estimation and control under adversarial attacks," *IEEE Transactions on Control of Network Systems*, 2020.

[68] S. Z. Yong, M. Zhu, and E. Frazzoli, "Switching and data injection attacks on stochastic cyber-physical systems: Modeling, resilient estimation, and attack mitigation," *ACM Transactions on Cyber-Physical Systems*, vol. 2, no. 2, pp. 1–2, 2018.

[69] L. F. Cómbita, J. Giraldo, A. A. Cárdenas, and N. Quijano, "Response and reconfiguration of cyber-physical control systems: A survey," in *IEEE Colombian Conference on Automatic Control (CCAC)*. IEEE, 2015, pp. 1–6.

[70] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[71] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[72] X. Huang, L. Wu, and Y. Ye, "A review on dimensionality reduction techniques," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 33, no. 10, p. 1950017, 2019.

[73] M. E. Valcher, "State observers for discrete-time linear systems with unknown inputs," *IEEE Transactions on Automatic Control*, vol. 44, no. 2, pp. 397–401, 1999.

[74] F. Blanchini and S. Miani, *Set-theoretic methods in control*. Springer, 2008.

[75] A. Chakrabarty, S. H. Żak, and S. Sundaram, "State and unknown input observers for discrete-time nonlinear systems," in *IEEE 55th Conference on Decision and Control (CDC)*. IEEE, 2016, pp. 7111–7116.

[76] B. Rosner, "Percentage points for a generalized esd many-outlier procedure," *Technometrics*, vol. 25, no. 2, pp. 165–172, 1983.

[77] J. Burkardt, "The truncated normal distribution," *Department of Scientific Computing Website, Florida State University*, vol. 1, p. 35, 2014.

[78] T. W. Anderson and I. Olkin, "Maximum-likelihood estimation of the parameters of a multivariate normal distribution," *Linear algebra and its applications*, vol. 70, pp. 147–171, 1985.

[79] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization." *Journal of machine learning research*, vol. 13, no. 2, 2012.

[80] A. J. Menezes, P. C. Van Oorschot, and S. A. Vanstone, *Handbook of applied cryptography*. CRC press, 2018.

[81] K. Gheitasi, M. Ghaderi, and W. Lucia, "A novel networked control scheme with safety guarantees for detection and mitigation of cyber-attacks," in *European Control Conference (ECC)*, 2019, pp. 1449–1454.

[82] C. Murguia, I. Shames, J. Ruths, and D. Nešić, "Security metrics and synthesis of secure control systems," *Automatica*, vol. 115, p. 108757, 2020.

[83] 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, 2000.

[84] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[85] A. Alanwar, Y. Stürz, and K. H. Johansson, "Robust data-driven predictive control using reachability analysis," *European Journal of Control*, p. 100666, 2022.

[86] M. Althoff, "On computing the minkowski difference of zonotopes," *arXiv preprint arXiv:1512.02794*, 2015.

[87] A. Matthias, "An introduction to cora 2015," in *Workshop on applied verification for continuous and hybrid systems*, 2015, pp. 120–151.

[88] A. Koch, J. Berberich, and F. Allgöwer, "Provably robust verification of dissipativity properties from data," *IEEE Trans. on Automatic Control*, vol. 67, no. 8, pp. 4248–4255, 2021.

[89] L. Yang and N. Ozay, "Scalable zonotopic under-approximation of backward reachable sets for uncertain linear systems," *IEEE Control Systems Letters*, vol. 6, pp. 1555–1560, 2021.

[90] N. Kochdumper and M. Althoff, "Representation of polytopes as polynomial zonotopes," *arXiv preprint arXiv:1910.07271*, 2019.

[91] Y. Chen and N. Ozay, "Data-driven computation of robust control invariant sets with concurrent model selection," *IEEE Transaction on Control Systems Technology*, vol. 30, no. 2, pp. 495–506, 2021.

[92] S. V. Rakovic, E. C. Kerrigan, K. I. Kouramas, and D. Q. Mayne, "Invariant approximations of the minimal robust positively invariant set," *IEEE Trans. on automatic control*, vol. 50, no. 3, pp. 406–410, 2005.

[93] M. Mejari, A. Gupta, and D. Piga, "Data-driven computation of robust invariant sets and gain-scheduled controllers for linear parameter-varying systems," *IEEE Control Systems Letters*, vol. 7, pp. 3355–3360, 2023.

[94] C. Escudero, C. Murguia, P. Massioni, and E. Zamaï, "Safety-preserving filters against stealthy sensor and actuator attacks," *IEEE Conference on Decision and Control*, pp. 5097–5104, 2023.

[95] D. Shi, Z. Guo, K. H. Johansson, and L. Shi, "Causality countermeasures for anomaly detection in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 63, no. 2, pp. 386–401, 2017.

[96] M. N. Kurt, Y. Yılmaz, and X. Wang, "Distributed quickest detection of cyber-attacks in smart grid," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 8, pp. 2015–2030, 2018.

[97] Y. Guan and X. Ge, "Distributed attack detection and secure estimation of networked cyber-physical systems against false data injection attacks and jamming attacks," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 4, no. 1, pp. 48–59, 2017.