SECURITY OF CONSTRAINED CYBER-PHYSICAL SYSTEMS

Kian Gheitasi

A thesis

in

The Department

of

Concordia Institute for Information Systems Engineering (CIISE)

Presented in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy

Concordia University

Montréal, Québec, Canada

January 2022

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: Kian Gheitasi

Entitled: 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 the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Chair
Chun-Yi Su

External Examiner
Justin Ruths

External to Program
Luis Rodrigues

Examiner
Arash Mohammadi

Examiner
Amr Youssef

Thesis Supervisor (s)
Walter Lucia

Approved by

Chair of Department or Graduate Program Director

Date of Defence

Dean,

Abstract

Security of constrained Cyber-Physical Systems

Kian Gheitasi, Ph.D.

Concordia University, 2022

In this thesis, the safety and security problems in Cyber-Physical Systems (CPSs) are addressed. In general, CPSs are referred to as physical systems tightly coupled with computation and communication capabilities, which have the potential to improve traditional engineering systems in terms of efficiency, reliability, and performance. However, such added features come along with potential vulnerabilities to cyber-attacks, as testified by the different types of cyber-attacks reported against CPSs. In the last decade, several control solutions have been proposed to detect such attacks and mitigate their impact on CPSs.

In the first part of this thesis, we show that most of the studied attacks, if performed for a finite-time duration, can be straightforwardly detected in the post-attack phase. Moreover, we show the existence of a new type of cyber-attacks, namely finite-time covert attacks, affecting both constrained and unconstrained control systems. It is formally proved that this class of attacks is undetectable, during their actions and after their termination, if the anomaly detector is implemented on the controller side of the CPS. To design such attacks against unconstrained control systems, we combine a finite impulse response receding-horizon filter and reachability arguments. On the other hand, for constrained control systems we resort to a Set-Theoretic Model Predictive Control (ST-MPC) approach leveraging robust reachability arguments.

In the second part of the thesis, we consider a constrained control system, subject to state and control input constraints, and we propose a novel networked control architecture to ensure the plant's safety, i.e., fulfillment of plant's safety constraints in the presence of cyber-attacks on the communication channels, regardless of attacker's actions and duration. To this end, two different detectors are proposed to detect attacks on the setpoint

signal as well as on the control inputs and sensor measurements. In addition, an Emergency Controller (EC), local to the plant, is designed to replace the networked controller whenever an attack is detected. The concept of a robust $N-$step attack-safe region is introduced to ensure that the EC is activated, regardless of the detector performance, at least one step before the safety constraints are violated.

In the third part of the thesis, we propose a novel networked control architecture aiming to minimize the tracking performance degradation under cyber-attacks. On the plant side, a local controller is designed to take care of attacks on the actuation channel. In particular, given a finite number of pre-determined admissible safe equilibrium points, this unit exploits a Voronoi partition of the state space and a family of dual-mode set-theoretic model predictive controllers to safely confine, in a finite number of steps, the system state into the closest robust control invariant region. On the other hand, on the controller side, the reference tracking controller operations are enhanced with an add-on module in charge of dealing with attack occurrences on the measurement channel. Specifically, by leveraging the Voronoi partition used on the plant's side and robust reachability arguments, the objective of this unit is to reduce the tracking performance loss by allowing a supervised system open-loop evolution until the best possible outcome in terms of tracking is achieved.

Acknowledgments

I would like to thank the following people, without whom I would not have been able to complete this research, and without whom I would not have made it through my Ph.D. degree!

First of all, I would like to express my deepest gratitude and appreciation to my supervisor, Dr. Walter Lucia, whose insight and knowledge into the subject matter steered me through this research. I also appreciate his enthusiasm for the project, his support, encouragement, and patience which provided me with the opportunity to walk on the right path.

During these four years, I got through hard times, which was not possible to pass without the support of my family. Words can't express how much I appreciate the love, support, and encouragement from my wife, Maryam. Also, I am grateful for the love and guidance of my parents and my brothers during all these years. This accomplishment would not have been possible without them.

I would also like to extend my deepest gratitude to all my fellow teammates, and friends at Concordia University, especially Maryam, Mohsen, Shima, and Amirreza, for their invaluable contribution, helpful advice, and unwavering support.

Last but not least, I am also grateful to my friends Saeid, Amir Moradi, Ehsan, Sepehr, Amir, Cristian, Antonello, Flavia, Niki, and Hadis for making my life so bright and colorful.

Contents

List of Figures

List of Tables

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
DoS:	Denial of Service
LTI:	Linear Time Invariant
RCI:	Robust Control Invariant
CIA:	Confidentiality, Integrity, and Availability
CSTR:	Continuous-Stirred Tank Reactor
IID:	Independent and Identically Distributed
LQG:	Linear Quadratic Gaussian
LQI:	Linear Quadratic Integral
FIR:	Finite Impulsive Response
RHUF:	Receding Horizon Unbiased FIR
DoA:	Domain of Attraction
T-DoA:	Tracking Domain of Attraction
UFTCA:	Undetectable Finite-Time Covert Attack
MPC:	Model Predictive Control
ST-MPC:	Set-Theoretic Model Predictive Control
QP:	Quadratic Programming
CG:	Command Governor
MN-D:	Monotonically Non-Decreasing

MN-I:	Monotonically Non-Increasing
EC:	Emergency Controller
SA:	Smart Actuator
SC:	Safety Controller
ROSC:	Robust One-Step Controllable
MAC:	Message Authentication Code

Chapter 1

Introduction and Literature Review

1.1 Cyber-Physical Systems

The term Cyber-Physical System (CPS) refers to physical systems equipped with communication and control capabilities. CPSs are widely used in our society from small to large-scale control systems, such as unmanned aerial vehicles, chemical plants, autonomous transportation systems, smart grids, and water distribution systems [1], [2]. They have the potential to improve traditional engineering systems in terms of efficiency, reliability, and performance. Nevertheless, improved capabilities come along with novel vulnerabilities to cyber-attacks targeting the cyber infrastructures and communication channels. In the control community, CPSs are typically abstracted as networked control systems where adversarial agents can affect the closed-loop system performance by performing cyber-attacks against the communication channels [3], [4]. The control community has been very active in studying the security, safety, and privacy issues associated with CPSs, see, e.g., [5–10], and references therein. Although extensive research has been done to address Networked Control Systems (NCS) problematic [11], the study of cyber-attacks affecting NCS is still a relatively young research area [10]. Cyber-attacks on industrial control systems can cause irrecoverable damages with huge financial and economic loss [12]. Several cyber-attacks affecting CPSs have been reported in the last decade, see e.g., the recent report on the

unauthorized access to the SCADA system at a US water treatment facilities [13], the recent attack against Florida water utilities [14], the Maroochi water breach [15], the well-known Stuxnet [16] and the Industroyer [17] malware. Due to economical and political interests, the number of cyber-threads on physical control systems has increased in recent years.

The networked control architecture in Figure. 1.1 is considered in this thesis, where C-P and P-C networks denote the communication channels between the plant and the control center (e.g., where the control logic, state estimator, and detection strategies are implemented) and the C-C channel is the communication channel between the command center (e.g., where the reference/setpoint signal is decided) and the control center.

Such a general architecture allows us to study the security and safety of different control architectures against different cyber-attacks. In some distributed control systems, the setpoint reference signal is not locally available to the controller but generated by a distributed command center. In such a setup, the attacker can alter the setpoint reference signal in order to prevent the plant from reaching the desired setpoint. Such a control scenario finds application in several domains such as the formation of unmanned vehicles [18] and smart grids [19], just to name a few. In other control system applications, the control center is networked, and the control input and sensor measurement signals are transmitted through the communication channels. In this setup, control inputs and sensor measurements signals might be subject to cyber-attacks. This setup applies to any control system where the plant is spatially distributed from the controller.

From a control point of view, to guarantee the security and safety of CPSs, different problems must be addressed [10]: i) attack detection strategies must be developed to discover the presence of cyber-attacks; ii) responsive emergency countermeasures must be designed to mitigate the attacker's actions while maintaining the overall system performance in a possibly degraded but acceptable level, e.g., preserving state and input constraints; iii) the control system must be capable of recovering normal operations and

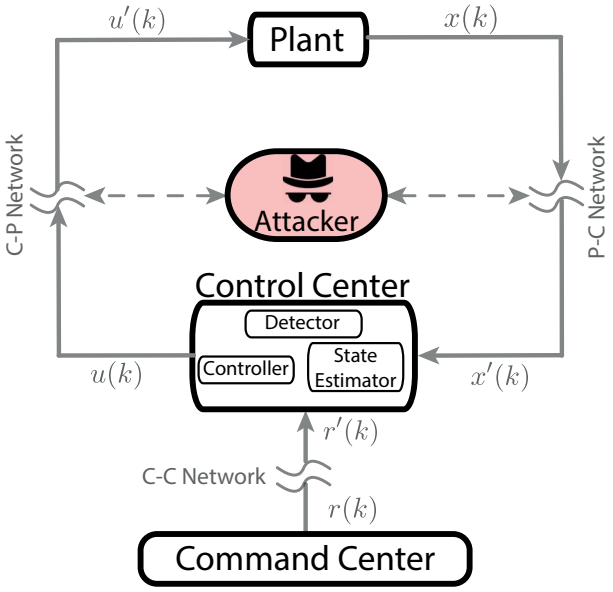


Figure 1.1: Considered Control Architecture

level of performance once the cyber-attack is ended.

1.2 Cyber-Attacks and Attack Detection Strategies

In order to be able to find appropriate solutions to the security problems highlighted in the previous section, the first step is to investigate different classes of attacks and their capabilities. In [20, 21], first, a 3-D attack classification is proposed to characterize the adversary's system knowledge, disclosure, and disruptive resources, then, different categories of cyber-attacks are defined accordingly. Two main classes of cyber-attacks against CPSs can be identified, Denial of Service (DoS) attacks and False Data Injection (FDI) attacks. DoS attacks prevent the transmitted data from reaching the destination, while FDI attacks alter the transmitted data to deceive the receiver. Typically, in the networked control systems, the problem of detection DoS attack is considered straightforward, especially if the communication channels are per se very reliable (as they typically are in SCADA system [22]) and the attack cannot be confused as poor network connectivity. On the other hand, the problem of detecting FDI is more challenging. Indeed, of particular

relevance are the classes of FDI attacks capable of affecting the control system performance while remaining stealthy (undetected) against standard passive residual-based detectors. Well-known examples of such attacks are zero-dynamics [23, 24], replay [25] and covert attacks [26]. In addition, new classes of undetectable attacks have been studied in [27–30], recently. In [27], it has been shown that local covert attacks can be performed with less disruptive resources w.r.t. traditional covert attacks. In [28], a covert channel technique has been presented to show that a compromised networked controller is able to leak private information to an eavesdropper who has access to the measurement channel. In [29], a robust pole-dynamics attack has been proposed against the control systems with unstable pole dynamics.

In the literature, several strategies have been proposed to diagnose malfunctions/faults in the control systems, see [31], [32]. However, such strategies, have been proved to be unable to reveal intelligent cyber-attacks [33]. As a consequence, different ad-hoc passive and active cyber-attack detection strategies have been developed, see [6,34] and references therein. Passive detection mechanisms are proved to be functional against DoS attacks, and simple FDI attacks on the actuation and measurement channels. In contrary, such detectors have limitations to detect specialized attacks such as covert, replay and zero-dynamics attacks [3], [35]. Therefore, active detection mechanisms are proposed to deal with these classes of attacks, see eg., [9, 35–43], and references therein. In [36] and [37], a moving target-based detector has been developed to detect covert attacks on the CPSs. In [38], a sensor coding strategy has been introduced to detect stealthy sensor attacks. The problem of detecting zero-dynamics and replay attacks has been investigated in [35] and [39, 40], respectively. In [9], a blended detection mechanism is proposed to deal with different types of attacks. In [41], centralized and distributed observer-based detection and identification strategies have been proposed to solve the problem of attack detection and identification for a large group of attacks. In [42], deep learning solutions are used to design attack detectors in the CPS context. Finally, in [43], the authors propose a

detection method to reveal attack vectors against nonlinear control systems. On the other hand, the problem of detecting setpoint attacks did not receive enough attention and it has only been recently investigated in [44, 45] where a detection strategy has been developed by taking advantage of the features of the command governor control strategy [46–48].

1.3 Safety and Security of CPSs

Responsive emergency countermeasures should be taken into account when an attack is detected, to ensure safety, security, and recovery of CPSs. To this end, several research studies have been performed on robust and resilient state estimation, see e.g., [49–53] and references therein. In [49], an l_0-based state estimator is proposed to ensure the resilience of the estimation in the presence of attacks. In [50], an observer with adaptive switching mechanism is proposed to reach an asymptotically stable observation error system under cyber-attacks, and in [52], 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 [53], a new state estimator in delta-domain is proposed against joint sensor and actuator attacks. As long as the design of resilient control strategies for CPSs under attacks is concerned, relevant are the contribution in [54–64] and reference therein. In [55], a variation of the receding-horizon control law is proposed to deal with the replay attacks. In [56], an active security control approach is presented for CPSs under DoS attacks. Similarly, in [57], optimal control strategies are developed using game theory in the delta domain to deal with DoS attacks, while in [58], the trade-off between system resilience and network bandwidth capacity is investigated. In [59] a mitigation approach against cyber-attacks is proposed by taking advantage of an improved adaptive resilient control scheme, and in [60], an attack-resilient receding-horizon control law is proposed to deal with replay attacks. In [61], a robust set-theoretic control paradigm is exploited for a constrained CPS to ensure that the plant can be recovered in a-priori known number of steps after an attack. In [62], an actuator security index is proposed to

protect the vulnerable actuators from cyber-attacks, and in [63], a data-based technique is used to learn the best defense strategy in the presence of replay attacks. Finally, in [64], a nonlinear encoding/decoding signal against integrity attacks has been proposed to detect anomalies and preserve the CPS's nominal performance, regardless of attacks.

1.4 Thesis Motivation and Contribution

First, to the best of our knowledge, existing studies have shown the existence of intelligent attacks that are undetectable, by design, during the attack actions. However, well-known classes of undetectable attacks such as covert attacks can be easily detected by any passive detector when the attack actions are terminated. Therefore, an important open question addressed in this thesis is related to the existence of a class of finite-time attacks that are undetectable, with respect to passive anomaly detectors [34] both when the attack is ongoing and afterward. Such a question is relevant in CPS applications where the attacker is interested in repeatedly or intermittently affecting the CPS performance without ever being detected. For example, in a modern water-treatment facility [22], or a power system [65], a malicious entity might be interested in stealing water/energy repeatedly (whenever it is needed), for a finite amount of time, and without ever triggering an anomaly. In this thesis, we show the existence of at-least a class of finite-time undetectable attacks, hereafter named finite-time covert attacks.

Secondly, the state-of-the-art lacks a solution to detect setpoint attacks and control solutions capable of preserving the plant's safety for constrained control systems regardless of the attack actions and durations. With respect to this problem, by considering constrained CPSs (e.g., CPSs subject to state and input constraints), this thesis proposes a control solution allowing the detection of different intelligent attacks while preserving the safety of the plant, regardless of the cyber-attack duration and actions.

Finally, the literature lacks control solutions aiming to deal with the reference tracking

control problem for constrained CPSs in the presence of cyber-attacks on both the actuation and measurement channels. In this regard, this thesis proposes a novel networked control architecture that aims to minimize the tracking performance degradation under cyber-attacks. Such a research can be considered a first attempt towards addressing the reference tracking problem in CPSs.

1.5 Thesis Layout

The main objectives of this research are twofold:

- Showing the existence of finite-time stealthy covert attacks.
- Designing a control architecture capable of:
  - Detecting cyber-attacks on the C-P, P-C, C-C 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 defined.
- In chapter 3, a new class of attacks, namely finite-time covert attacks, is designed against both constrained and unconstrained control systems. It is formally proved that this class of attacks remains stealthy during the attack and after its termination.
- In chapter 4, a novel networked control architecture is designed to ensure plant's safety against a variety of cyber-attacks that can affect the communication channels in cyber-physical systems.

- In chapter 5, the reference tracking control problems for constrained CPSs under attacks is investigated and a control strategy aiming to minimize the performance degradations under attacks is proposed.

- Finally, in chapter 6, conclusions about the performed research studies are given, and possible research directions are discussed.

1.6 Publications related to the thesis

- **Kian Gheitasi**, and Walter Lucia. "A worst-case approach to safety and reference tracking for cyber-physical systems under network attacks", IEEE Transactions on Automatic Control, 2021. (under review)

- [66] **Kian Gheitasi**, and Walter Lucia. "Undetectable Finite-Time Covert Attack on Constrained Cyber-Physical Systems", IEEE Transactions on Control of Network Systems, 2021.

- [67] **Kian Gheitasi**, and Walter Lucia. "A safety preserving control architecture for cyber-physical systems", International Journal of Robust and Nonlinear Control, 2021.

- [30] **Kian Gheitasi**, and Walter Lucia. "A Finite-Time Stealthy Covert Attack Against Cyber-Physical Systems", International Conference on Control, Decision and Information Technologies (CoDIT), 2020.

- [68] **Kian Gheitasi**, Mohsen Ghaderi, and Walter Lucia. "A novel networked control scheme with safety guarantees for detection and mitigation of cyber-attacks", European Control Conference (ECC), 2019.

- [44] Walter Lucia, **Kian Gheitasi**, and Mohsen Ghaderi. "Setpoint Attack Detection in Cyber-Physical Systems", IEEE Transactions on Automatic Control, 2020.

- [9] Mohsen Ghaderi, **Kian Gheitasi**, and Walter Lucia. "A blended active detection strategy for false data injection attacks in cyber-physical systems", IEEE Transactions on Control of Network Systems, 2020.
- [45] Walter Lucia, **Kian Gheitasi**, and Mohsen Ghaderi. "A command governor based approach for detection of setpoint attacks in constrained cyber-physical systems", 2018 IEEE Conference on Decision and Control (CDC).
- [69] Mohsen Ghaderi, **Kian Gheitasi**, and Walter Lucia. "A novel control architecture for the detection of false data injection attacks in networked control systems", 2019 American Control Conference (ACC).

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:

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) + \omega(k) \\ y(k) &= Cx(k) + \nu(k) \end{aligned} \tag{2.1}$$

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, and $u(k) \in \mathbb{R}^m$ is the control input vector.

In the rest of this thesis, two variants of the model (2.1) are used:

- Stochastic model: in this model $\omega(k)$, and $\nu(k)$ are process and output noises obtained from independent and identically distributed (IID) normal distributions with zero-mean and covariances $\mathcal{Q} > 0$ and $\mathcal{R} > 0$, respectively, i.e., $\omega(k) \sim \mathcal{N}(0, \mathcal{Q})$, and $\nu(k) \sim \mathcal{N}(0, \mathcal{R})$.

- Robust model: in this model $\omega(k)$, and $\nu(k)$ are bounded disturbances, where

$$\begin{aligned} \omega(k) \in \mathcal{D}_x \subset \mathbb{R}^n, \quad 0_n \in \mathcal{D}_x \\ \nu(k) \in \mathcal{D}_y \subset \mathbb{R}^n, \quad 0_n \in \mathcal{D}_y \end{aligned} \tag{2.2}$$

  Moreover, set-membership state and input constraints are prescribed as follows:

$$u(k) \in \mathcal{U}, \quad x(k) \in \mathcal{X} \quad \forall\, k \in \mathbb{Z}_+ \tag{2.3}$$

  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.

Under attacks on the C-P and P-C channels, the plant dynamics becomes:

$$\begin{aligned} x(k+1) &= Ax(k) + Bu'(k) + \omega(k) \\ y(k) &= Cx(k) + \nu(k) \end{aligned} \tag{2.4}$$

where $u'(k)$ is the control signal received by the plant.

2.1.2 Control Center

In this section, the tracking controller, state estimator, and the anomaly detector subsystems in Figure. 1.1, are introduced.

2.1.2.1 Controller

By denoting with $x_c(k) \in \mathbb{R}^{n_c}$ the state of the controller, its actions can be generically described as

$$u(k) = f(x_c(k), y(k), r(k)) \tag{2.5}$$

where $r(k)$ is the desired reference signal and $f(\cdot,\cdot,\cdot)$ is a stabilizing control logic. If the robust model (2.2) is of interest, we assume that the control logic (2.5), in the absence of attacks, fulfills the constraints (2.3) despite any disturbance realization (2.2). Moreover, its Domain of Attraction (DoA) is $\mathcal{X}_c \subseteq \mathcal{X}$ [70].

2.1.2.2 State Estimator

By considering the stochastic model (2.1), the steady-state Kalman predictor [71] is:

$$\hat{x}(k+1) \;\; = \;\; A\hat{x}(k) + Bu(k) + L(y'(k) - C\hat{x}(k)) \tag{2.6}$$

where $\hat{x}(k)$ is the state estimation and $y'(k)$ is the measurement vector received by the state estimator module. By assuming (A, C) detectable and (A, B_q), $B_q^T B_q = Q$ stabilizable, then $L = APC^T(CPC^T+R)^{-1}$ is the steady-state Kalman gain with $P^T = P > 0$ obtained as $P = \lim_{k\to\infty} P(k)$ where

$$P(k+1) = AP(k)A^T + Q - AP(k)C^T(CP(k)C^T + R)^{-1}CP(k)A^T \tag{2.7}$$

is the Riccati equation that, for $k \to \infty$, admits only one positive semidefinite solution.

2.1.2.3 Anomaly Detector

By using the Kalman filter (2.6)-(2.7) the residual signal

$$res(k) = y'(k) - C\hat{x}(k) \tag{2.8}$$

is an Independent and Identically Distributed (IID) Gaussian process with zero-mean and covariance $\Sigma = CPC^T + \mathcal{R}$, i.e., $res(k) \in \mathcal{N}(0, \Sigma)$. Therefore, an anomaly detector can be designed by online checking the statistical properties of $res(k)$. In principle, the following binary hypothesis test can be performed:

- $\mathcal{H}_0$ (normal operations / no attack), if $\begin{cases} E[res(k)] = 0 \\ E[res(k)res^T(k)] = \Sigma \end{cases}$
- $\mathcal{H}_1$ (anomaly / cyber-attack), if $\begin{cases} E[res(k)] \neq 0 \\ E[res(k)res^T(k)] \neq \Sigma \end{cases}$

In practice, different approximations of the above test have been proposed and, in the sequel, the well-established χ^2 test is used

$$g(k) = \sum_{i=k-M+1}^{k} res(i)^T \Sigma^{-1} res(i) \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\lessgtr}} \tau \tag{2.9}$$

where $M > 0$ is the detection window size and $\tau > 0$ is a threshold value that can be analytically tuned to obtain the desired probability of false alarms, see e.g., [72].

2.2 Definitions

Definition 2.1. *(**Stealthy attack**) A cyber-attack is stealthy if it can reduce the control system performance while remaining undetected for an indefinitely large time interval [73].* □

Definition 2.2. *(**Safe Plants**) The plant (2.1) is considered safe with guaranteed performance recovery if : (i) under attacks, no constraints (2.3) violations occur, and (ii) after the attacks, the attack-free control performance can be recovered [67].*

Definition 2.3. *(**False Data Injection Attack**) Let's consider two subsystems S_1 and S_2 and 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 $v^a(k)$, [74], i.e.,

$$v'(k) = v(k) + v^a(k) \tag{2.10}$$

Definition 2.4. *(**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).*

Definition 2.5. *(**Minkowski/Pontryagin set sum and difference**) Given two sets $\mathcal{A} \subset \mathbb{R}^n$ and $\mathcal{B} \subset \mathbb{R}^n$, the Minkowski/Pontryagin set sum and difference are defined as follows:*

$$\mathcal{A} \oplus \mathcal{B} := \{a + b | a \in \mathcal{A}, b \in \mathcal{B}\}$$
$$\mathcal{A} \ominus \mathcal{B} := \{a \in \mathcal{A} | a + b \in \mathcal{A}, \forall b \in \mathcal{B}\}$$

Definition 2.6. *(**Robust Control Invariant set**) A set $\mathcal{S} \subseteq \mathcal{X}$ is said Robust Control Invariant (RCI) [75] for (2.1) under (2.2)-(2.3) if $\forall x \in \mathcal{S}, \exists u \in \mathcal{U} : Ax + Bu + \omega \in \mathcal{S}, \ \forall \omega \in \mathcal{D}_x$.* □

2.3 Cyber-attacks

In this section, different cyber-attacks on the C-P and P-C channels are classified according to the available resources.

2.3.1 Attacker's resources

In CPSs, a communication channel is considered secure against cyber-attacks if the *Confidentiality*, *Integrity* and *Availability* properties (also known as the CIA triad) cannot be compromised by the attacker [76]. Definition of each property in CPS is as follows:

- **Confidentiality:** Refers to the ability to keep information secret from unauthorized users. If the adversary cannot obtain/read information/data related to system by

eavesdropping on the communication channels, the NCS has confidentiality.

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

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

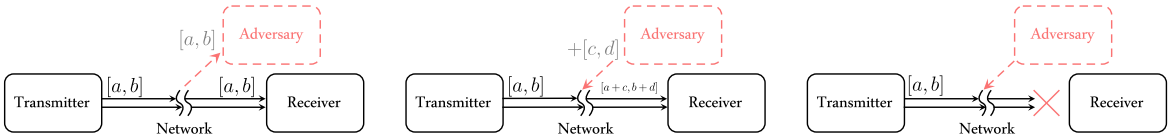


Figure 2.1: Lack of (2.1a) Confidentiality. (2.1b) Integrity (2.1c) Availability

In general, the more information and resources that the attacker has, the more complex attacks can be performed. According to [20], cyber-attacks can be classified based to their available resources: (i) model knowledge, (ii) disclosure resources and (iii) disruptive resources (see Figure. 2.2).

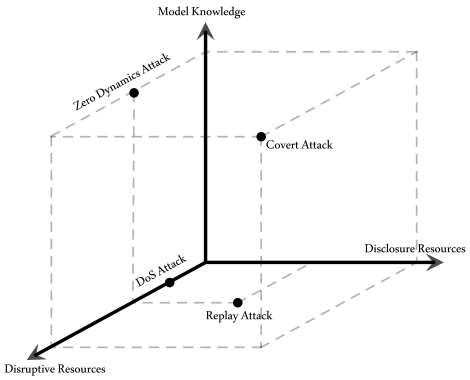


Figure 2.2: Cyber-Physical attack space [21]

- **Disclosure resources:** is the set of communication channels where the attacker can violate the confidentiality property (e.g., 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 (e.g., 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.

2.3.2 Classes of attacks

In this section, the most important classes of cyber-attacks are reviewed.

2.3.2.1 Denial of Service (DoS) attack

In a DOS attack, the adversary tries to prevent sensor measurements or control inputs from reaching the control center or the plant, respectively. Therefore, such an attack breaks the feedback loop, forcing the system to evolve in an open-loop fashion. Due to its behavior, DoS attacks could be mis-recognized with the poor network connection. Performing DoS attacks does not need any system model knowledge and disclosure resources (see Figure.2.2). Usually, this attack is being performed by jamming the communication channels, so only disruptive resources are needed to launch such an attack.

2.3.2.2 Replay attack

In a replay attack, 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 from the plant. Performing replay attack needs disclosure resources on measurement channels and disruptive resources on both channels. This attack scenario is performed in two phases (see Figure.2.4):

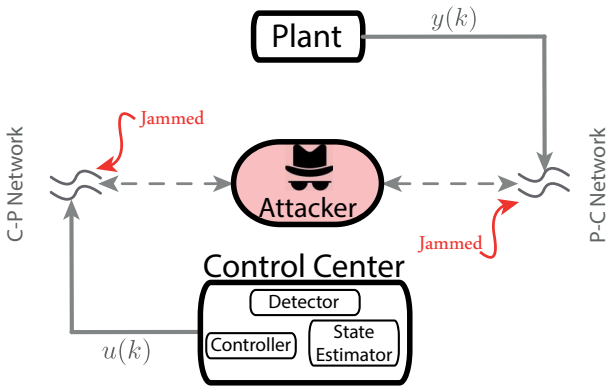


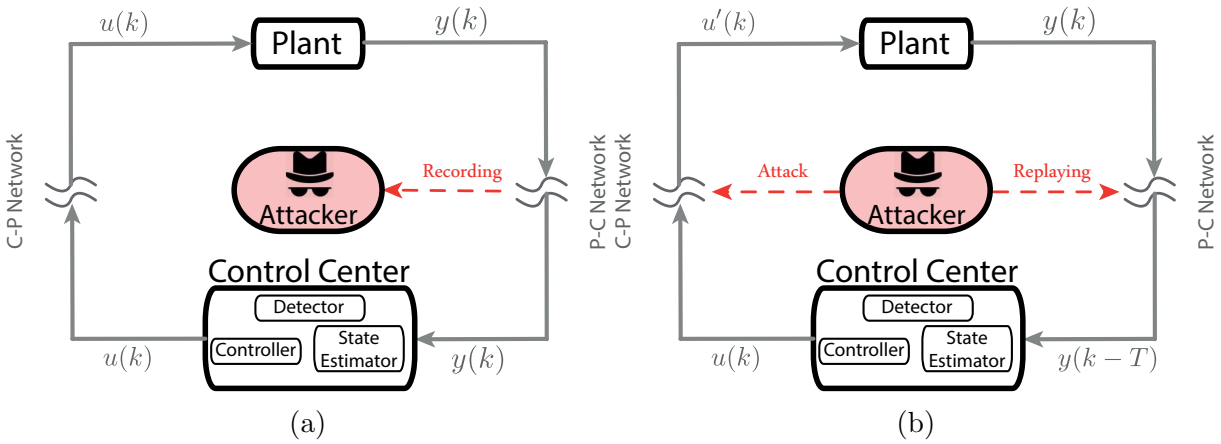
Figure 2.3: DoS attack on CPS

- **Phase I. Recording:** The adversary records the measurement signals for a period of time (attack duration).

$$y_{rec}(\bar{k}) = y(k'); \quad \bar{k} = 1 : T; \quad k - T \leq k' < k \tag{2.11}$$

- **Phase II. Replaying:** The recorded data are replayed while malicious inputs are injected in the actuation channel.

$$y'(k') = y_{rec}(\bar{k}); \quad \bar{k} = 1 : T; \quad k \leq k' < k + T \tag{2.12}$$



(a) (b)

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

2.3.2.3 Zero-dynamics attack

The aim of a zero-dynamics attack (Figure. 2.5a) is to excite the unstable zero(s) of the system (2.1), to reduce control performance while producing zero output. Zero-dynamics attacks can be performed if the attacker is aware of the plant's model (2.1) (including knowledge about the initial state of the plant), and it possesses disruptive resources on the actuation channel (see Figure. 2.2).

A zero-dynamics attack is a particular FDI attack on the actuation channel where the attack vector is computed as follows:

$$\begin{aligned} u^a(k) &= \lambda^k g \\ u'(k) &= u(k) + u^a(k) \end{aligned} \tag{2.13}$$

where λ are the non-minimum phase zeros of the system that cause the following matrix to lose rank:

$$\begin{bmatrix} \lambda I - A & -B \\ C & D \end{bmatrix}$$

and $g \neq 0$ is the input zero direction for the chosen zero obtained by solving:

$$\begin{bmatrix} \lambda I - A & -B \\ C & D \end{bmatrix} \begin{bmatrix} x(0) \\ g \end{bmatrix} = 0. \tag{2.14}$$

2.3.2.4 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^a into the C-P channel:

$$u'(k) = u(k) + u^a(k) \tag{2.15}$$

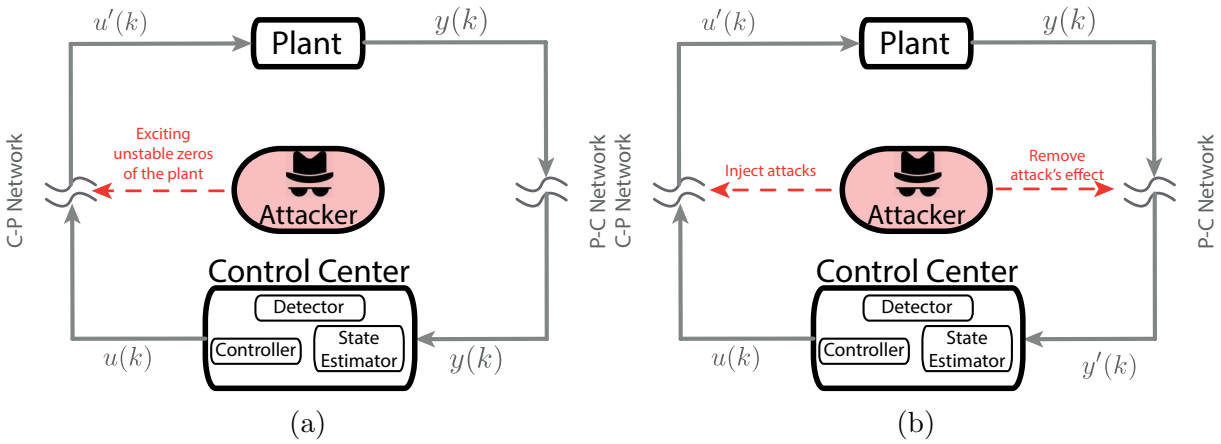


Figure 2.5: (2.5a) Zero-dynamics attack. (2.5b) Covert attack

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:

$$y'(k) = y(k) - y^a(k) \tag{2.16}$$

where $y^a(k)$ is the effect of attack $u^a(k)$ on the plant dynamics, such that:

$$y^a(k) = C \sum_{j=0}^{k-1} (A^j B u^a(k-1-j)). \tag{2.17}$$

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

2.4 Detection mechanisms

In what follows, different detection mechanisms that have been proposed to deal with the previously introduced attacks, are presented.

2.4.1 Watermarking approach

The use of watermarked control inputs has been proposed in [77], [39] to detect stealthy steady-state replay attacks on the measurement channel. The main idea of this active mechanism is to inject private randomly generated perturbations into the control signal to reveal the presence of attacks:

$$u(k) = u^*(k) + \Delta u(k) \tag{2.18}$$

where $u^*(k)$ is assumed to be the desired control input. In [25], it has been shown that if $u^*(k)$ is designed to minimize the following LQG regulation cost:

$$J = \min_{u^*} \lim_{T\to\infty} \frac{1}{T+1} E[\sum_{k=0}^{T-1} x^T(k) W x(k) + u^{*T}(k) U u^*(k)] \tag{2.19}$$

where W, and U are positive semidefinite and definite cost matrices, respectively, then, the control action (2.18) introduces a degradation $\Delta u(k)$ which is a Gaussian distribution with the covariance $\mathscr{Q}$. The LQG performance after adding the authentication signal $\Delta u(k)$ is given by:

$$J' = J + \Delta J \tag{2.20}$$

where J is the optimal performance of the controller and ΔJ is the deviation from the optimal solution, caused by the injection of Δu:

$$\Delta J = trace[(U + B^T S B)\mathscr{Q}] \tag{2.21}$$

and S satisfies the following Riccati equation:

$$S = A^T S A + W - A^T S B (B^T S B + U)^{-1} B^T S A. \tag{2.22}$$

It has been proved in [25] that the watermarking signal (2.18) allows the χ^2 detector (2.9) to reveal the presence of the replay attacks.

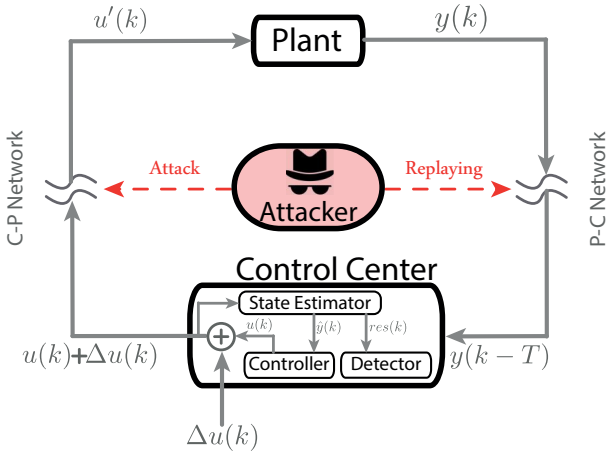


Figure 2.6: Watermarking approach

2.4.2 Sensor coding approach

The main idea of the sensor coding mechanism, proposed in [38], [78], is to encode the sensor measurement data in a way that the adversary cannot inject false data that are undetectable by a χ^2 anomaly detector. As a consequence, it prevents the existence of stealthy covert attacks by keeping the real measurements signal secret from the attacker.

To remain stealthy, an attacker aims to increase the state estimation error, while keeping the residual signal of the anomaly detector (2.8), at a small level. By coding the sensor measurement signal, the residual signal will be a function of the coding algorithm. Therefore, no attacks can remain stealthy (i.e., control the value of the residual signal to avoid detection) without any knowledge about the coding algorithm.

2.4.3 Moving target approach

The idea behind the moving target solution, developed in [37], [36] is to introduce further auxiliary plant dynamics, unknown to the attacker, that behave as a moving target

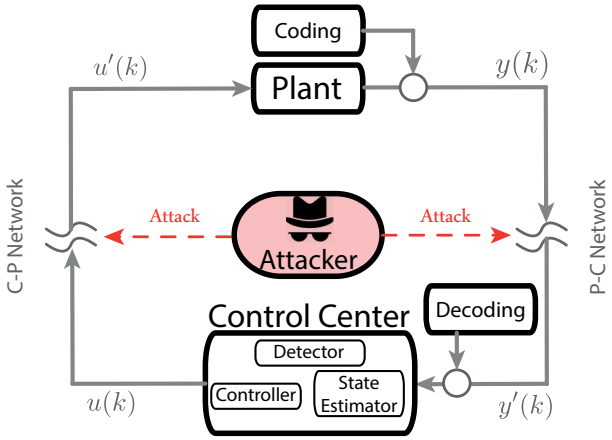


Figure 2.7: Sensor coding approach

mechanism. As proposed in [36], the extra dynamics have the following structure:

$$\begin{aligned} \tilde{x}(k+1) &= A_1(k)\tilde{x}(k) + A_2(k)x(k) + B_1u(k) + \tilde{\omega}(k) \\ \tilde{y}(k) &= C_1(k)\tilde{x}(k) + \tilde{\nu}(k) \end{aligned} \tag{2.23}$$

where $\tilde{x}$ is the state of the extended system, $\tilde{y}$ is the output of the extended system, $A_1(k), A_2(k), B_1(k)$, and $C_1(k)$ are IID random matrices which are independent of the sensor and process noises. By coupling the auxiliary dynamics with the plant model (2.1), the augmented/extended plant dynamics are

$$\begin{aligned} \begin{bmatrix} \tilde{x}(k+1) \\ x(k+1) \end{bmatrix} &= \mathcal{A}(k) \begin{bmatrix} \tilde{x}(k) \\ x(k) \end{bmatrix} + \mathcal{B}(k)u(k) + \begin{bmatrix} \tilde{\omega}(k) \\ \omega(k) \end{bmatrix} \\ \begin{bmatrix} \tilde{y}(k+1) \\ y(k+1) \end{bmatrix} &= \mathcal{C}(k) \begin{bmatrix} \tilde{x}(k+1) \\ x(k+1) \end{bmatrix} + \begin{bmatrix} \tilde{\nu}(k) \\ \nu(k) \end{bmatrix} \end{aligned} \tag{2.24}$$

where:

$$\mathcal{A}(k) \triangleq \begin{bmatrix} A_1(k) & A_2(k) \\ 0 & A \end{bmatrix}, \quad \mathcal{B}(k) \triangleq \begin{bmatrix} B_1(k) \\ B \end{bmatrix}, \quad \mathcal{C}(k) \triangleq \begin{bmatrix} C_1(k) & 0 \\ 0 & C \end{bmatrix}, \tag{2.25}$$

As shown in [37], if the auxiliary extended dynamics are kept secret from the attacker (e.g., by generating them from a pseudo-random generator initialized with a secret seed shared between the plant and the controller), then the covert attack (2.15)-(2.17) cannot be performed for the simple reason that the attacker cannot exactly compute $y^a(k)$ as in (2.17).

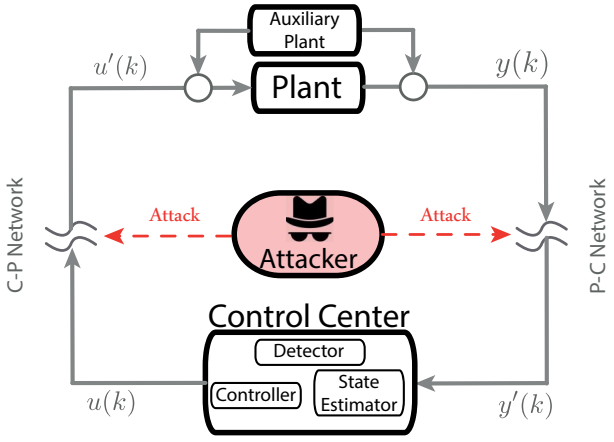


Figure 2.8: Moving target approach

2.5 Set-Theoretic Model Predictive Control

In this section, by considering robust plant model dynamics, the basic ST-MPC strategy [79] is presented, and its main properties are summarized.

Consider an equilibrium pair (x_{eq}, u_{eq}) for the plant dynamics (2.1) and the system's state and control input constraints (2.3). The objective of the control strategy is to steer, in a finite-number of steps, the state trajectory $x(k)$ into a neighborhood of x_{eq}, regardless of disturbance (2.2) realization while fulfilling the state and input constraints (refer to the equation). To this end, first, the ST-MPC strategy is offline designed according to the following steps [75]:

- **Offline-step 1:** A terminal static state feedback controller for the unconstrained and disturbance-free model-dynamics in (2.1) is designed in order to asymptotically

drive the state trajectory of the system $x(k)$ into x_{eq},

$$u(k) := -K(x(k) - x_{eq}) + u_{eq}, \tag{2.26}$$

where K is the controller gain.

- **Offline-step 2:** By considering the state and control input constraints (2.3), the smallest RCI region around x_{eq} (see Definition 2.6), namely $\mathcal{T}_0$, is computed and associated with the terminal controller in (2.26). This RCI region is computed according to [81] considering the following requirements:

$$\mathcal{T}_0 \subseteq \mathcal{X}, \quad u(k) \in \mathcal{U}; \quad \forall k \tag{2.27}$$

- **Offline-step 3:** The controller designed in the previous steps has the ability to keep all the states within the terminal region $\mathcal{T}_0$, regardless of any admissible disturbance realization. The Domain of Attraction (DoA) of the terminal controller can be enlarged by computing a family of robust one-step controllable sets, namely $\{\mathcal{T}_i\}_{i=1}^{N}$, using the following recursive definition:

$$\begin{aligned} \mathcal{T}_i \;:=\; & \{x \in \mathcal{X} : \exists u \in \mathcal{U}, \forall \omega \in \mathcal{D}_x, \text{ s.t. } Ax + Bu + \omega \in \mathcal{T}_{i-1}\} \\ & \{x \in \mathcal{X} : \exists u \in \mathcal{U}, \text{ s.t. } Ax + Bu \in \tilde{\mathcal{T}}_{i-1}\} \end{aligned} \tag{2.28}$$

where $\tilde{\mathcal{T}}_i := \mathcal{T}_i \ominus \mathcal{D}_x$, and N is the number of calculated sets. The recursion stops when the DoA of the designed controller, which is the union of the calculated sets in (2.28), i.e., $\bigcup_{i=0}^{N} \mathcal{T}_i$, saturates or it covers the set of all admissible initial conditions, e.g., $\mathcal{X} \subseteq \bigcup_{i=0}^{N} \mathcal{T}_i$. Please note that the above one-step controllable sets can be numerically computed in Matlab, using, e.g., MPT3 toolbox [80].

Online, the family of robust one-step controllable sets is exploited to calculate, the control input $u(k)$ that steers the state trajectory of the plant into the terminal region $\mathcal{T}_0$.

- **Online-step 1:** Find the smallest set $i(k)$ that contains $x(k)$, i.e.

$$i(k) := \min\{0 \leq i \leq N : x(k) \in \mathcal{T}_i\} \tag{2.29}$$

- **Online-step 2:** If $i(k) = 0$, i.e., the state is in the terminal region, then apply the terminal controller. Otherwise, solve the following optimization problem to compute a control input imposing that the one-step evolution of the system belongs to $\mathcal{T}_{i(k)-1}$.

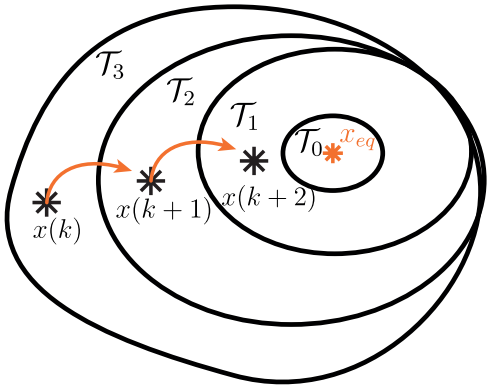


Figure 2.9: Family of robust one-step controllable sets, and the state trajectory evolution

$$\begin{aligned} u(k) \quad = \quad & \arg\min_u J(x(k), u), \quad \text{s.t.} \\ & Ax(k) + Bu \in \tilde{\mathcal{T}}_{i(k)-1}, \\ & u \in \mathcal{U} \end{aligned} \tag{2.30}$$

where the cost function $J(x(k), u)$ can be arbitrary chosen to penalize any desired convex combination of control effort and convergence rate.

The aforementioned algorithm ensures that from any $x(0) \in \bigcup_{i=0}^{N} \mathcal{T}_i$, the plant state trajectory is uniformly ultimately bounded into the terminal region $\mathcal{T}_0$ in, at-most, N steps (see Figure. 2.9). Also, in the disturbance-free scenario, the state trajectory will asymptotically converge to x_{eq}.

Chapter 3

A Finite-Time Stealthy Covert Attack Against Cyber-Physical Systems

3.1 Introduction

In the last decade, several cyber-attacks against Cyber-Physical Systems (CPSs) have been investigated, and different active and passive control solutions have been proposed to assure the absence of undetectable attacks. In this chapter, we show that most of the studied attacks, if performed for a finite-time duration, can be straightforwardly detected in the post-attack phase. Also, given a proper justification for the existence of finite-time attacks, we show that a finite-time stealthy covert attack can be performed if the attacker takes ad-hoc actions before terminating. We propose a practical implementation of a finite-time covert attack on both constrained and unconstrained control systems. First, by combining a finite impulse response receding-horizon filter and reachability arguments, we design such attacks against unconstrained control systems. A simulation example, involving a quadruple-tanks water system, is shown to better clarify the capabilities of the designed attack. Then, we face a similar design problem but in a more challenging

setup (for the attacker) where the plant is subject to bounded but unknown disturbances and state and input constraints. In particular, by resorting to a set-theoretic control framework [75] and robust controllability arguments for constrained systems, we show that, under proper conditions, a finite-time stealthy covert attack exists. Moreover, for a given attack duration and attack objective (e.g., state configuration to reach under attack), we characterize the subspace of states from which the proposed attack is guaranteed to be successful. A simulation example of a Continuous-Stirred Tank Reactor (CSTR) system is provided to testify the proposed design effectiveness.

3.1.1 Contribution of the work

To the best of our knowledge, there are no works focusing on existence of finite-time stealthy attacks. Motivational examples for this study can be found in different domains. For instance, in energy or water distribution systems, see e.g., [22,82], an attacker might be interested in repeatedly/intermittently launching a finite-time undetectable attack to steal water or energy for a fixed amount of time. In such contests, undetectability is also desired in the post-attack phase. In this chapter, we show the existence of a particular class of undetectable finite-time attacks, namely finite-time covert attacks, that are undetectable during the attack actions and after their termination. Their existence is shown for both constrained and unconstrained control systems. Moreover, for constrained systems, the state-space region from which the attack is doable is characterized.

3.2 Finite-time stealthy attack against unconstrained control systems

In this section, we design a finite-time stealthy attack on unconstrained control systems.

3.2.1 Networked Control System

Let us consider the following discrete-time linear system

$$\begin{aligned} x(k+1) &= Ax(k) + Bu'(k) + \omega(k) \\ y(k) &= Cx(k) + \nu(k) \end{aligned} \tag{3.1}$$

where $k \in \mathbb{Z}_+ := \{0, 1, ...\}$, $x(k) \in \mathbb{R}^n$ is the plant state vector, $y(k) \in \mathbb{R}^p$ is the plant output vector, and $u'(k) \in \mathbb{R}^m$ is the received control input. Moreover, A, B and C are the system matrices of suitable dimensions, while $\omega(k)$ and $v(k)$ are the process and measurement noises obtained from independent and identically distributed (IID) normal distributions with zero-mean and covariances $\mathcal{Q} > 0$ and $\mathcal{R} > 0$, respectively, i.e., $\omega(k) \sim \mathcal{N}(0, \mathcal{Q})$, and $\nu(k) \sim \mathcal{N}(0, \mathcal{R})$.

Assumption 3.1. *We assume that the plant (3.1) is detectable and stabilizable.* □

3.2.1.1 Control Center

We assume that within the control center, a Linear Quadratic Gaussian (LQG) controller is used. It consists of a Kalman state estimator and an optimal LQ controller, i.e.,

$$u(k) = K(x(k) - x_{eq}) + u_{eq} \tag{3.2}$$

where K is the optimal LQ gain and (x_{eq}, u_{eq}) is an equilibrium pair for (3.1). Moreover, we assume that a χ^2-based anomaly detector is used to detect the presence of attacks. The Kalman filter, and the χ^2- detector are the ones introduced in section 2.1.2.2 and 2.1.2.3.

3.2.2 Attacker Model

Assumption 3.2. *We consider networked cyber-attacks capable of performing fine-time deception attacks in the plant-to-controller and controller-to-plant channels. In particular,*

by denoting with $\underline{k}^a \geq 0$, *and* $\overline{k}^a < \infty$ *the attack starting and ending time instants, the following class of attacks is defined [21]:*

- *Disclosure Resources - The attacker is capable of reading the signal* $u(k)$ *and* $y(k)$, $\forall\ \underline{k}^a \leq k \leq \overline{k}^a$

- *Disruptive Resources - The attacker can perform the following additive False Data Injection attacks (FDI), i.e.,*

$$\begin{aligned} u'(k) &= u(k) + u^a(k) \\ y'(k) &= y(k) + y^a(k) \end{aligned} \quad ,\forall\ \underline{k}^a \leq k \leq \overline{k}^a \tag{3.3}$$

  where $u^a(k) \in \mathbb{R}^m$ *and* $y^a(k) \in \mathbb{R}^p$ *are the injected attack vectors.*

- *Plant Knowledge - The attacker has perfect knowledge of the plant dynamical model (3.1), but he is not aware of the control center operations (e.g., controller, state-estimator and anomaly detector)*

Definition 3.1. *(Finite-Time Stealthy Attack) An attack on a CPS is said a finite-time attack if the attack can be performed only for a finite period of time, i.e.,* $\underline{k}^a \leq k \leq \overline{k}^a$. *A finite-time attack is said stealthy if it is capable of arbitrarily altering the closed-loop plant (3.1)-(3.2) performance for* $\underline{k}^a \leq k \leq \overline{k}^a$ *while remaining undetectable* $\forall\, k \geq \underline{k}^a$. □

3.2.3 Problem Formulation

Let's consider the networked control system illustrated in Figure. 1.1, described in the previous section. With the given setup and resources (see *Assumption* 3.2), it is possible to design advanced FDI attacks that bypass passive detection mechanisms (e.g., the χ^2 detector (2.9)) during their actions. Among others, notable examples of undetectable attacks are replay [25] and covert attacks [26].

It is possible to show, without resorting to standard technicalities, that a passive detection mechanism can effectively detect Replay or Covert attacks when the preventive

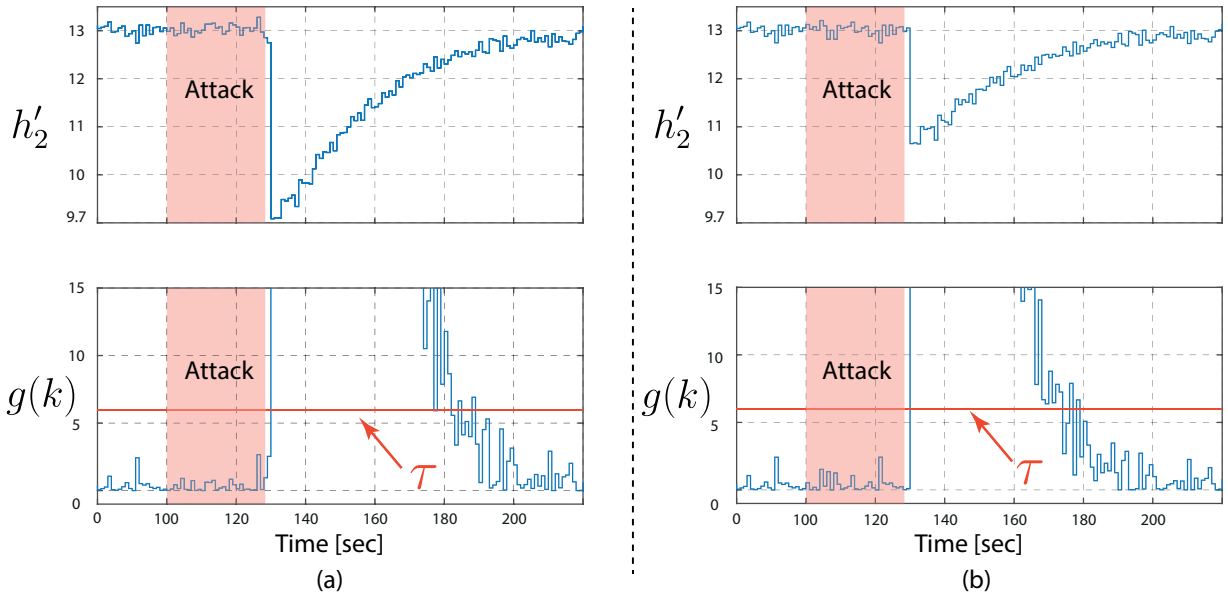


Figure 3.1: Replay (subplots (a)) and Covert (subplots (b)) attacks detection in the post-attack phase: corrupted sensor measurement h_2' for the water's level in tank 2 and χ^2 test.

measures, used to avoid detection, are terminated. To better explain this concept, we can refer to Figure. 3.1 where we have emulated the above attacks on a quadruple-tanks water system (please refer to the simulation section for the system description). In particular, the designed attacks start at $k = 100$ and end at $k = 130$. Since both attacks strategies are, by design, undetectable, then the residual signals, during the attack phase, do not reveal any anomaly. On the other hand, when the attack is terminated, an abrupt change of the sensor measurements and residual signals can be clearly appreciated. As a consequence, the passive detector (2.9) fails to detect the attacks during their actions, but it still reveals an anomaly in the post-attack phase.

The objective of this section is to show how it is possible to design a finite-time stealthy attack satisfying *Definition* 3.1. The problem can be stated as follows:

(CH3.2-O1) - Design of Finite-Time Stealthy Attacks: Let's consider the plant model (3.1), the networked control architecture (3.2), (2.6)-(2.9), and the set of attack's resources in *Assumption* 3.2. We want to show the existence of *Finite-Time Stealthy Attacks* (see *Definition* 3.1) capable of arbitrarily altering the plant state trajectory and

being undetectable, $\forall k$, by any detection strategy located only in the control center.

3.2.4 Finite-Time Stealthy Covert Attack Design

In what follows, a finite-time covert attack is designed. The section starts revising the standard covert attack operations, then it proceeds to highlight the design challenges for a finite-time covert-attack realization and proposing a solution which combines a deadbeat receding horizon Kalman FIR filter [83] and reachability arguments. Finally, under a minimum-time attack duration requirement, it is proved that the proposed covert-attack is capable of arbitrarily altering the system performance while remaining undetectable to any detector located in the control center.

3.2.5 Covert-Attack [26]

Under the presence of the FDI attack (3.3), we can rewrite, for linearity, the output evolution of the signal $y'(k)$, as the sum of three distinct contributions:

$$y'(k) = (y^u(k) + y^{u^a}(k)) + y^a(k) \tag{3.4}$$

with

$$y^u(k) := CA^k x_0 + C\sum_{j=0}^{k-1}(A^j(Bu(k-1-j) + \omega(k-1-j))) + \nu(k) \tag{3.5}$$

$$y^{u^a}(k) := C\sum_{j=0}^{k-1}(A^j B u^a(k-1-j)) \tag{3.6}$$

where $y^u(k)$ is system output evolution in the absence of attacks and $y^{u^a}(k)$ is the effect of an input attack $u^a(k)$ on the output measurements. Therefore, given the assumed attack resources (*Assumption* 3.2), an attacker can perform a covert attack, undetectable for any passive detector [26], if $\forall\, k$:

(i) The attacker injects an arbitrary input signal $u^a(k)$

(ii) The attacker cancels out the input attack effect $y^{u^a}(k)$ from the output signal by injecting $y^a(k) = -y^{u^a}(k)$

Remark 3.1. *In addition to (i)-(ii), if the plant's (3.1) state is unknown to the attacker, a preliminary state-estimation phase might be needed to reconstruct an accurate estimation of the system's state, namely $\hat{x}^a(k)$. The latter enables the attacker to launch a covert attack capable of steering the state trajectory into any desired state configuration.*

Let's now consider a finite-time covert attack starting at $k = \underline{k}^a$ and ending at $k = \overline{k}^a$. To launch the attack (i)-(ii) for $\underline{k}^a \leq k \leq \overline{k}^a$ and ensure undetectability for $k > \overline{k}^a$ irrespective of the detection strategy used in the control center, the following aspects must be considered

- Before the FDI attack is started, an accurate state estimation $\hat{x}^a(k)$ must be obtained in a finite number of steps;
- Before the attack is terminated, the signal $y^{u^a}(k)$ must be dragged to zero, to ensure stealthiness in the post-attack phase, i.e., $y^{u^a}(k) \equiv 0$, $\forall\, k > \overline{k}^a$.

To take care of the above concerns, the 3-phase finite-time cover attack illustrated in Figure. 3.2 is designed.

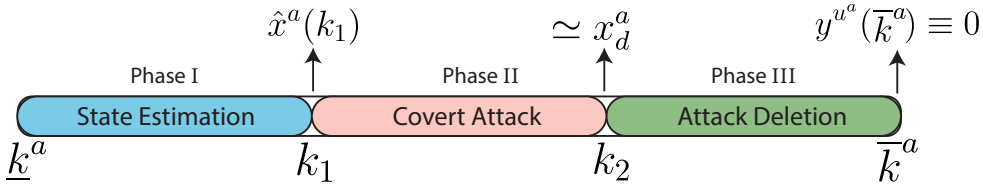


Figure 3.2: Attacker Strategy

3.2.6 Phase I - Finite-time state estimation

In Phase I, the attacker wants to obtain an unbiased estimation of the state of the system at $k = k_1$, i.e.,

$$\hat{x}^a(k_1) = E[x(k)] \tag{3.7}$$

where $E[\cdot]$ denotes the expected value. The attacker reads the transmitted sensor measurements $y(k)$ and command inputs $u(k)$ intercepted for $\underline{k}^a \leq k < k_1$.

A Finite Impulsive Response (FIR) filter must be used to obtain, in a finite number of time-steps k_1, an unbiased state-estimation $\hat{x}^a(k-1)$. In the sequel, we use the Receding Horizon Unbiased FIR (RHUF) proposed in [83] which batch implementation over a prediction horizon $N > 0$ with $n \leq N \leq k_1$ is:

$$\hat{x}^a(k) = H_B(Y(k-1) - \bar{B}_N U(k-1)) \tag{3.8}$$

where

$$\begin{aligned} H_B &= (\bar{C}_N^T E_N^{-1} \bar{C}_N)^{-1} \bar{C}_N^T E_N^{-1} \\ Y(k-1) &= [y(k-N)^T,\, y(k-N+1)^T, \ldots, y(k-1)^T]^T \\ U(k-1) &= [u(k-N)^T,\, u(k-N+1)^T, \ldots, u(k-1)^T]^T, \end{aligned} \tag{3.9}$$

and $\bar{C}_N$, $\bar{B}_N$, E_N are obtained from the following recursive definitions for $1 \leq i \leq N$:

$$\begin{aligned} \bar{C}_i &= \begin{bmatrix} \bar{C}_{i-1} \\ C \end{bmatrix} A^{-1} \\ \bar{B}_i &= \begin{bmatrix} \bar{B}_{i-1} & -\bar{C}_{i-1} A^{-1} B \\ 0 & -CA^{-1}B \end{bmatrix} \\ E_i &= \begin{bmatrix} E_{i-1} & 0 \\ 0 & R \end{bmatrix} + \begin{bmatrix} \bar{C}_{i-1} \\ C \end{bmatrix} A^{-1} Q A^{-T} \begin{bmatrix} \bar{C}_{i-1} \\ C \end{bmatrix}^T. \end{aligned} \tag{3.10}$$

Assumption 3.3. *The matrix A is assumed to be nonsingular for the discretized model in (3.1).*

The main RHUF state-estimator (3.7) properties can be summarized as follows (see [84] for formal proofs):

- If $N \geq n$ then the state estimation is unbiased, see (3.7), and it minimizes the covariance of the estimation error.

- No prior statistics information about the horizon initial state $x(k-N)$ are needed
- In a noise-free scenario, the state-estimator (3.7) has deadbeat property, and $\hat{x}^a(k) = x(k)$.

Remark 3.2. *The above properties show that an attacker can obtain an unbiased estimation of the system's state if $k_1 - \underline{k}^a \geq n$. Moreover, since the state-estimation error covariance decreases by increasing the prediction horizon, the attacker can design k_1 to achieve the desired estimation performance.*

3.2.7 Phase II - Finite-time covert actions

In Phase II, for $k_1 \leq k < k_2$, the attacker aims to steer the plant states towards a desired configuration x_d^a. To this end, the covert attack actions (i)-(ii) detailed in Section 3.2.5 are specialized as follows:

1. Solve the reachability problem

$$\begin{aligned} U_a^{'} &= \arg \min_{[u'(k_1),\ldots,u'(k_2-1)]} ||U'_{[k_1,k_2)}||_2^2, \ s.t. \\ & \quad z^a = R_{[k_1,k_2)} U'_{[k_1,k_2)} \end{aligned} \tag{3.11}$$

   where $R_{[k_1,k_2)} = \left[A^{k_2-k_1+1} \ \ldots, \ AB, \ B\right]$ is the reachability matrix and

$$z^a = x_d^a - A^{k_2-k_1+1}\hat{x}^a(k-1), \ \ U'_{[k_1,k_2)} = \begin{bmatrix} u'(k_1) \\ \vdots \\ u'(k_2-1) \end{bmatrix}$$

2. $\forall\, k_1 \leq k < k_2$

   - Inject the FDI

$$u^a(k) = u'(k) - u(k) \tag{3.12}$$

with $u'(k)$ obtained from (3.11).

- Inject the FDI $y^a(k) = -y^{u^a}(k)$ (see (3.6)).

Remark 3.3. *Since the plant (3.1) is assumed to be controllable, then the reachability problem (3.11) always admits a solution as long as the Phase II duration is bigger than n steps, i.e., $k_2 - k_1 \geq n$. Moreover, the optimization (3.11), picks, among all the admissible solution $U'_{[k_1,k_2)}$, the one at minimum energy, namely U'_a.*

It is also important to remark that the section aims only to show the existence of an attack capable of arbitrarily affecting the plant state-trajectory in a finite number of steps. For the sake of clarity, it is important to mention that only in the noise-free case it is ensured that $x(k_2) \equiv x_d^a$. Otherwise, depending on the noise realizations, x_d^a might not be exactly reached. This arises from the fact that the defined attack is an open-loop attack. Such drawback can be mitigated, if needed, by designing an attack vector $u^a(k)$ through a robust state-feedback tracking controller based on the RHUF filter.

3.2.8 Phase III - Finite-time attack deletion

In Phase III, for $k_2 \leq k \leq \overline{k}^a$, the attacker, before terminating its actions, aims to delete the attack effect on the system in order to avoid post-attack detection (see Figure. 3.1).

As previously shown for the output signal $y(k)$ (see (3.4)), we can exploit the superposition principle to separate the input attack effect in the state vector evolution, i.e.,

$$x'(k) = x^u(k) + x^{u^a}x(k) \tag{3.13}$$

with

$$\begin{aligned} x^u(k) &:= A^k x(0) + \sum_{j=0}^{k-1}(A^j(Bu(k-1-j) + \omega(k-1-j))) \\ x^{u^a}(k) &:= \sum_{j=0}^{k-1}(A^j B u^a(k-1-j)) \end{aligned} \tag{3.14}$$

where $x^{u^a}(k)$ is the state evolution under the effect of the input attack $u^a(k)$. The latter

allows us to express the reachability problem that consists of a set of attacker's actions $u^a(k)$ capable of dragging the vector $x^{u^a}(k)$ to 0_n. In particular, the reachability problem and the attacker's actions are the followings:

1. Solve the reachability deletion problem

$$\begin{aligned} U_d &= \arg \min_{[u'(k_2),\dots,u'(\overline{k}^a-1)]} ||U_{[k_2,\overline{k}^a)}||_2^2, \ s.t. \\ & z^d = R_{[k_2,\overline{k}^a)} U_{[k_2,\overline{k}^a)} \end{aligned} \tag{3.15}$$

where $R_{[k_2,\overline{k}^a)} = \left[A^{\overline{k}^a-k_2+1} \ \dots, \ AB, \ B\right]$ is the reachability matrix and

$$z^d = 0_n - A^{\overline{k}^a-k_2+1} x^{u^a}(k_2), \ \ U_{[k_2,\overline{k}^a)} = \begin{bmatrix} u^a(k_2) \\ \vdots \\ u(\overline{k}^a-1) \end{bmatrix}$$

2. $\forall\, k_2 \leq k < \overline{k}^a$

   - Inject the FDI $u^a(k)$ computed in (3.15).
   - Inject the FDI $y^a(k) = -y^{u^a}(k)$ (see (3.6)) in the measurements channel.

Remark 3.4. *As previously commented in Remark 3.3, the reachability problem (3.15) has a solution as long as the Phase III duration is bigger than n steps, i. e.* $\overline{k}^a - k_2 \geq n$. *Moreover, from linearity, the attack deletion, i.e.,* $x^{u^a} \equiv 0, \forall\, k \geq \overline{k}^a$, *is irrespective of the noises realizations.*

3.2.9 Finite-time covert attack - Properties

Proposition 3.1. *Let's consider the plant model (3.1), the residual signal (2.8) and the attack model (3.3). If the attack duration is equal or greater to* $3n$, *i.e.,* $\overline{k}^a - \underline{k}^a \geq 3n$, *then the 3-phase finite-time covert attack designed in (3.7)-(3.15) is capable of (i) steering the*

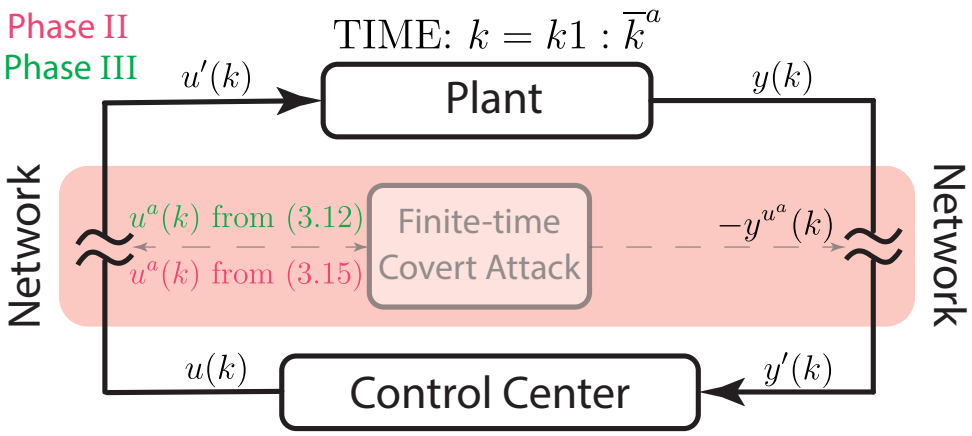


Figure 3.3: Phase II and III: Covert attack and attack deletion

state-trajectory $x(k)$ *towards any desired state-space configuration* $x_d^a \in \mathbb{R}^n$ $\forall\, \underline{k}^a \leq k \leq \overline{k}^a$ *while remaining undetectable* $\forall\, k \geq \underline{k}^a$ *to any detector strategy only based on the residual signal (2.8).*

Proof. The result (i) can be proved by collecting the result in Section 3.2.7. As long as the attack stealthiness is of interest (ii), it is possible to notice that during Phase II and III, i.e., $k_1 \leq k \leq \overline{k}^a$, the attack is undetectable because the FDI on the sensor measurements are designed, according to the covert paradigm [26], to completely cancel out the effect of FDI on the input channel, see (3.6). As a consequence, the residual signal (2.8) will be completely unaffected by input attack, and the attack detection probability is unchanged irrespective of the used detection rule. Moreover, Phase III has been designed to completely nullify the state evolution due to the attack in Phase II. As a consequence, by construction, for $k > \overline{k}^a$, the residual signal is still unchanged ensuring stealthiness in the post-attack phase. Finally, since each stage of the designed attack requires $n-$steps to guarantee feasibility, $3n$ represents the minimum time-duration needed for an attacker to launch a finite-time attack satisfying the objectives in **(CH3.2-O1)**. □

Remark 3.5. *From the results in Proposition 3.1 it is possible to provide two key requirements for the design of detection and mitigation strategies against the designed finite-time covert attacks. In particular:*

(R1) *Finite-time covert attacks, like standard covert-attacks [26], cannot be detected by detection mechanisms only located in the control center;*

(R2) *Any effective detection and mitigation strategy must be able to detect and mitigate the attack in less than $2n$ steps. Indeed, $2n$ is the minimum time needed for the attacker to complete Phases I and II and steering the plant's trajectory in any hazardous configuration.*

Finally, we would like to mention, that existing active detection mechanisms for covert attacks such as moving target [37, 69], can be in principle used to detect finite-time covert attacks. Nevertheless, to make such a strategy effective, proper modifications are required to ensure the timing requirements stated in (R2).

3.2.10 Simulation Example

In this section, the quadruple-tanks water system described in [85] is used as a test-bed example for the designed finite-time covert attack.

The system consists of four tanks interconnected as shown in Figure. 3.4, for which the levels of water, $x = [h_1, \dots h_4]^T$, define the system's state variables. The system is regulated by two valves which receive commands in terms of voltage levels (v_1 and v_2), i.e., $u = [v_1, v_2]^T$. The measured variables are the levels of water in tanks 1 and 2, i.e., $y = [h_1, h_2]^T$. The nonlinear system dynamics have been linearized around the following operating equilibrium point [85]

$$x_{eq} = [12.4,\ 12.7,\ 1.8,\ 1.4]^T,\ \ u_{eq} = [3,\ 3]^T$$

and discretized using a sampling time $T_s = 1\,\text{sec}$. The following matrices A, B, C are taken

from [85], and the initial condition of the plant states are $x = \begin{bmatrix} 11.9 & 12.4 & 1.6 & 0.9 \end{bmatrix}^T$.

$$A = \begin{bmatrix} 0.975 & 0 & 0.042 & 0 \\ 0 & 0.977 & 0 & 0.044 \\ 0 & 0 & 0.958 & 0 \\ 0 & 0 & 0 & 0.956 \end{bmatrix}$$

$$B = \begin{bmatrix} 0.0515 & 0.0016 \\ 0.0019 & 0.0447 \\ 0 & 0.0737 \\ 0.0850 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} 0.2 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 \end{bmatrix} \tag{3.16}$$

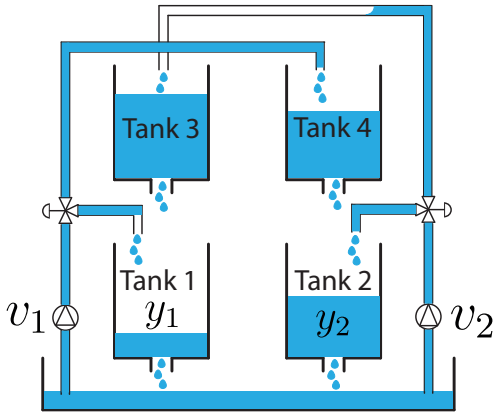


Figure 3.4: Quadruple-Tanks Water System

The control center has been configured to use an LQ controller, a Kalman filter state-estimator and a χ^2 anomaly detector with $M = 5$, and $\tau = 19.92$ obtained for a false alarm rate equal to 3%.

We assume, as in Figure. 1.1, that the control center is networked and that the attacker takes complete control of measurements and actuation channels for $96 \leq k \leq 136$ [sec]. The attacker aims to steal water from the tanks 1 and 2 by reaching the water overflow level. We assume that the water overflow occurs if $h_1 > 12.7$ or $h_2 > 13$.

The finite-time covert attack is designed as follows:

- Phase I ($96 \leq k < 100\,[\text{sec}]$): The attacker reads $u(k)$ and $y(k)$ and uses the RHUF filter (3.8) with a horizon $N = 4$ to obtain $\hat{x}^a(100)$.

- Phase II ($100 \leq k < 118\,[\text{sec}]$): The attacker, to steal water from thank 1 and 2, imposes the overflow state configuration $x_d^a = [12.8,\ 13.3,\ 2.25,\ 1.55]^T$, via the optimization (3.11), customized as follows:

$$\begin{aligned} U_a' &= \arg \min_{[u'(100),\ldots,u'(118)]} ||U'_{[100,118)}||_2^2,\ s.t. \\ & [A^{17}B,\ A^{16}B,\ \ldots,\ B]U'_{[100,118)} = x_d^a - A^{17}\hat{x}^a(100) \end{aligned}$$

- Phase III ($118 \leq k \leq 136\,[\text{sec}]$): The attacker, to cancel out the attack effect, performs the deletion attack (3.15) to bring the water levels, in the tanks 1 and 2, back to the expected values.

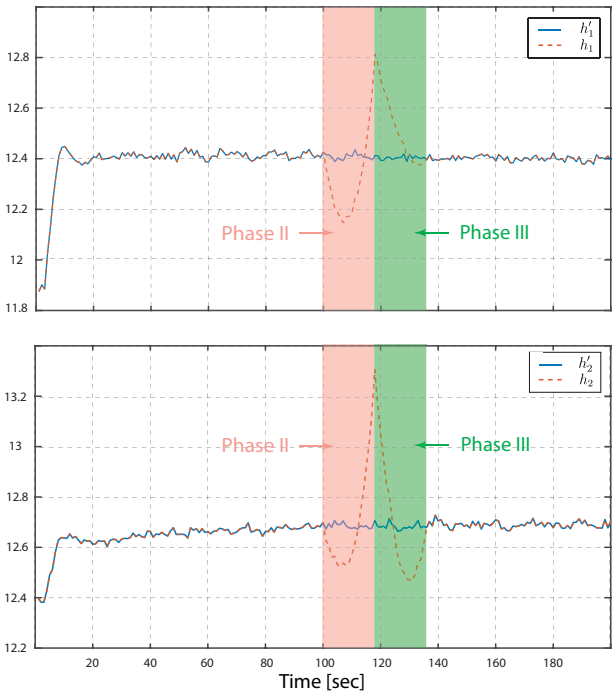


Figure 3.5: Sensor measurements: real (h_1, h_2) an and corrupted (h_1', h_2')

The main simulation results, related to phases II and III are collected in Figures. 3.5-3.7. In Figure. 3.5 the actual, h_1, h_2 and corrupted, h_1', h_2' sensor measurements are

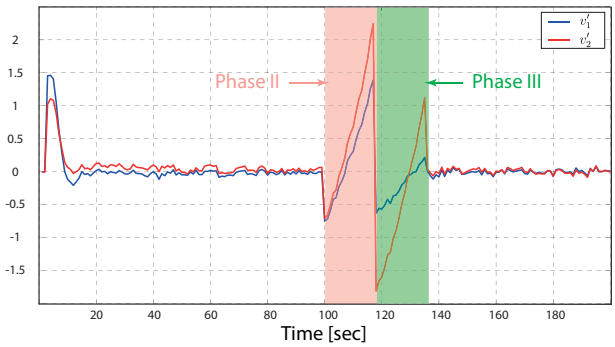


Figure 3.6: Corrupted Input signals $u' = [v_1', v_2']^T$

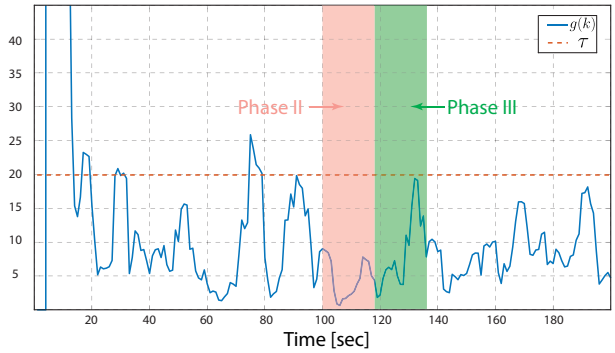


Figure 3.7: χ^2 test with an expected false alarm rate $\leq 3\%$.

reported for the tanks 1 and 2. It is possible to appreciate that, in phase II, due to FDI attack on the control signals (see Figure. 3.6) the water levels h_1 and h_2 reach, in a finite number of steps, the overflow condition ($h_1 > 12.7$ or $h_2 > 13$). Indeed, at $k = 118\,[\text{sec}]$, $h_1 = 12.8$ and $h_2 = 13.3$. On the other hand, due to the covert FDI attack on the sensor measurement, the signal received by the Control Center are unaffected by the attack actions. As a consequence, the χ^2 detector shown in Figure. 3.7 does not reveal the presence of the attack. In phase III, similar arguments can be given. The attacker, before leaving the system, injects proper inputs to bring back the water levels in tank 1 and tank 2 to the controller's expected value. In particular, at $k = 136\,[\text{sec}]$, $h_1 = 12.4$ and $h_2 = 12.7$, see Figure. 3.5. Moreover, in Figures. 3.5 and 3.6, it is possible to notice that at the end of phase III, i.e., at $k = 136\,[\text{sec}]$, the corrupted input signal $u'(k)$ and corrupted sensor measurements $y'(k)$ match the expected input and output. The latter guarantees that in the post-attack phase, i.e., $k > 136\,[\text{sec}]$, the residual signal is unaffected by the

previously performed attack. In particular, χ^2 test shown in Figure. 3.7 maintains a false alarm rate $\leq 3\%$ both during and after the attack.

3.3 Finite-time stealthy attack against constrained control systems

In this section, we design a finite-time stealthy attack on constrained control systems.

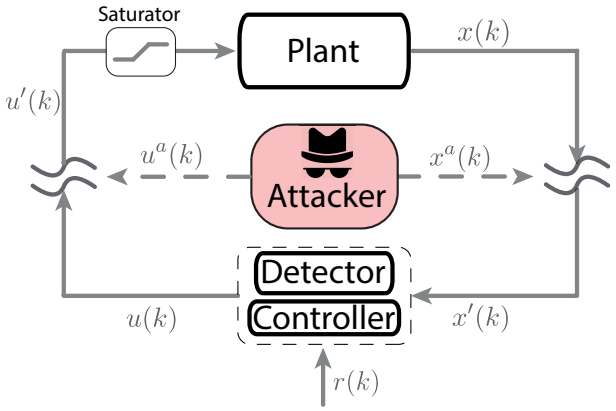


Figure 3.8: Networked Control Architecture

3.3.1 Networked Control System Setup

Consider the following discrete-time linear system

$$x(k+1) = Ax(k) + Bu'(k) + B_d d(k) \tag{3.17}$$

where the index $k \in \mathbb{Z}_+ = \{0, 1, ...\}$ denotes discrete-time instants, $x(k) \in \mathbb{R}^n$ the vector of the states, $u'(k) \in \mathbb{R}^m$ the control input vector received by the plant, $d(k) \in \mathbb{R}^d$ a bounded unknown disturbance, and A, B and B_d are the system matrices with appropriate dimensions. The unknown disturbance $d(k)$ is such that

$$d(k) \in \mathcal{D} \subset \mathbb{R}^d, \quad 0_d \in \mathcal{D} \tag{3.18}$$

with $\mathcal{D}$ a compact set. The actuators' physical limitations impose the following saturation constraint on $u'(k)$

$$u'(k) \in \mathcal{U} \subset \mathbb{R}^m, \quad 0_m \in \mathcal{U} \tag{3.19}$$

with $\mathcal{U}$ a compact set, while the state are desired to be constrained into the set

$$x(k) \in \mathcal{X} \subset \mathbb{R}^n, \quad 0_n \in \mathcal{X} \tag{3.20}$$

where $\mathcal{X}$ is a compact set.

Assumption 3.4. *We assume that the plant (3.17) is stabilizable.*

3.3.1.1 Control Center

3.3.1.1.1 Controller: The networked controller is a tracking state-feedback controller designed to comply with the constraints (3.19)-(3.20) despite the disturbance realization (3.18). By denoting with $x_c(k) \in \mathbb{R}^{n_c}$ the state of the controller, its actions are generically described as

$$u(k) = f(x_c(k), x(k), r(k)) \tag{3.21}$$

where $r(k)$ is the desired reference signal and $f(\cdot, \cdot, \cdot)$ the control logic. In what follows, we assume that the control logic (3.21) is given and its Domain of Attraction (DoA) is $\mathcal{X}$.

3.3.1.1.2 Detector: A dynamic passive anomaly detector, leveraging the received state measurements $\{x'(k)\}$ and computed control inputs $\{u(k)\}$, is used to reveal anomalies/ cyber-attacks, see [34] for a survey paper. Without loss of generality, the anomaly detection rule can be described as

$$\text{anomaly}(k) = \Phi(\{x'(t)\}_{t=0}^{k}, \{u(t)\}_{t=0}^{k-1}, \mathcal{D}) \tag{3.22}$$

where $\Phi(\cdot,\cdot,\cdot)$ is the binary attack detection logic. Moreover, anomaly$(k) = 1$ if an attack is detected, 0 otherwise.

3.3.2 Attacker Model

The attacker is capable of corrupting the communication channels between the plant and the controller. In particular, the three-dimensional characterization of the attack is [21]:

- The attacker is aware of the plant model (3.17);
- The attacker can read the control signal $u(k)$ and the state measurement vector $x(k)$.
- The attacker can produce a deception attack to change the control signal ($u(k) \rightarrow u'(k) \in \mathbb{R}^m$) and state measurements ($x(k) \rightarrow x'(k) \in \mathbb{R}^n$) received by the plant and the networked controller, respectively.

Given the available resources, and a desired target state $x_d \in \mathbb{R}^n$, the attacker is able to compute (e.g., by resorting to the method described in [81]) an admissible small RCI target set $\mathcal{X}_d \subset \mathbb{R}^n$, centered in x_d.

3.3.3 Problem Formulation

In this section, the existence and design of finite-time covert attacks, undetectable in the post-attack phase, are investigated. The problem of interest can be formulated as follows:

Undetectable Finite-Time Covert Attack (UFTCA): Consider the networked control system shown in Figure. 3.8, and the target RCI region $\mathcal{X}_d \subset \mathbb{R}^n$ centered in $x_d \in \mathbb{R}^n$. Design a finite-time deception attack of duration $\bar{T} \in \mathbb{Z}_+$, i.e.,

$$\begin{gathered} u'(k) = u(k) + u^a(k), \quad x'(k) = x(k) + x^a(k) \\ \underline{k} \leq k \leq \bar{k}, \;\; \bar{k} - \underline{k} = \bar{T} \end{gathered} \tag{3.23}$$

with $u^a(k) \in \mathbb{R}^m$ and $x^a(k) \in \mathbb{R}^n$ arbitrarily FDI vectors, such that:

- **(CH3.3-O1)** The attack is capable of steering the state trajectory within the target set $\mathcal{X}_d$ for at least one time instant, i.e., $x(k) \in \mathcal{X}_d$, $\forall\, k \in [k_{in},\, k_{out}]$, where $k_{in} \geq \underline{k}$, $k_{out} \leq \bar{k}$, and $k_{out} - k_{in} \geq 0$.

- **(CH3.3-O2)** Regardless of the used dynamic detector (3.22), the attack does not trigger any alarm during its actions ($\underline{k} \leq k \leq \bar{k}$) and afterward ($k > \bar{k}$).

3.3.4 Basic Covert Attack

In this section, the detectability in the post-attack phase of the covert attack, introduced in [26], is discussed. Under the presence of FDI attacks on the control signal (i.e., $u^a(k) \neq 0$), the system (3.17) evolves as:

$$x(k+1) = Ax(k) + B(u(k) + u^a(k)) + B_d d(k) \tag{3.24}$$

For linearity, it is possible to write

$$x(k) = x^u(k) + x^{u^a}(k) \tag{3.25}$$

where

$$x^u(k) = A^k x(0) + \sum_{j=0}^{k-1} A^j \left(Bu(k-1-j) + B_d d(k-1-j)\right) \tag{3.26}$$

$$x^{u^a}(k) = \sum_{j=0}^{k-1} A^j B u^a(k-1-j) \tag{3.27}$$

Notice that $x^u(k)$ denotes the state evolution of the system due to the initial condition, control input, and disturbance realization, while $x^{u^a}(k)$ is the state evolution of the system due to the presence of the input attack vector $u^a(k)$.

According to the covert attack introduced in [86], an attacker can arbitrarily affect the state trajectory (3.24) while remaining undetected by (3.22) if

- $$u'(k) = u^a(k) + u(k) \in \mathcal{U},\ \forall k \text{ s.t. } \underline{k} \leq k \leq \bar{k} - 1 \tag{3.28}$$

- $$x^a(k) = -x^{u^a}(k),\ \forall k \text{ s.t. } \underline{k} + 1 \leq k \leq \bar{k} \tag{3.29}$$

Remark 3.6. *The above attack is, by construction, undetectable for* $\underline{k} \leq k \leq \bar{k}$ *irrespective of* Φ *used in (3.22) [26]. Visually, by referring to Figure. 3.9a, regardless of* $u^a(k) \neq 0$*, the system state received by the controller is always equal to the expected one, i.e.,* $x'(k) = x^u(k),\ \forall\, \underline{k} \leq k \leq \bar{k}$*.*

On the other hand, when the covert attack is terminated (i.e., $k > \bar{k}$*) we have that*

$$\begin{aligned} x(k) &= x^u(k) + A^{k-\bar{k}} x^{u^a}(\bar{k}) \\ x'(k) &= x(k), \qquad\qquad k > \bar{k} \end{aligned}$$

Therefore if $x^{u^a}(\bar{k}) \neq 0_n$*, then for some* $k > \bar{k}$*,* $x'(k) \neq x^u(k)$ *and such discrepancy can be leveraged by (3.22) to detect an anomaly (see Figure. 3.9b). In conclusion, attack stealthiness in the post-attack phase (i.e.,* $\forall\, k > \bar{k}$*) is guaranteed irrespectively of the detector logic and any disturbance realization if* $x^{u^a}(\bar{k}) = 0_n$*.* □

(a) $\underline{k} \leq k \leq \bar{k}$ (b) $k > \bar{k}$

Figure 3.9: State mismatch during and after the attack.

3.3.5 *UFTCA* design

In this section, we design a finite-time covert attack fulfilling the objectives **(CH3.3-O1)**-**(CH3.3-O2)** stated in the *UFTCA* problem formulation.

First, the challenges of such a design are highlighted:

1. The attack must determine a control sequence $\{u^a(k)\}_{k=\underline{k}}^{\bar{k}-1}$ where $\exists\, k \in [\underline{k},\, \bar{k})$ such that the state trajectory enters the desired RCI region $\mathcal{X}_d$ for at least one time instant. Moreover, the control actions must fulfill the input saturation constraint (3.19) and be robust against any admissible disturbance realization (3.18) and controller (3.21) actions.

2. The attacker, to avoid any possibility of post-attack detection, must make sure that $x^{u^a}(\bar{k}) = 0_n$ (see the analysis in Remark 3.6). Such an objective must be robust against disturbance realization (3.18) and controller (3.21) actions.

3. Given a finite amount of time $\bar{T}$, the attacker must be able to determine (before starting the attack), the set of initial state conditions $\mathcal{X}_a \subseteq \mathcal{X}$, from which the attack is guaranteed to succeed.

Given the constrained and uncertain nature of the above problem, here we provide a solution, based on a robust set-theoretic model predictive control (ST-MPC) paradigm [75, 79, 87]. Please note that other MPC paradigms or constrained control strategies can, in principle, be used instead of ST-MPC. Such a choice is mainly motivated by the fact that ST-MPC will allow to offline define the controller's domain of attraction (union of robust one-step controllable sets) and the worst-case number of steps required to robustly reach the attacker's objectives.

By resorting to a *divide et impera* approach, the UFTCA design problem is divided in two phases (see Figure. 3.10). In the first phase, a covert attacks is designed to ensure that $\forall\, x(k) \in \mathcal{X}$, there exists a sequence of attack actions $\{u^a(k)\}$ such that the state

trajectory is driven within $\mathcal{X}_d$. In the second phase, an attack deletion strategy is designed to ensure that $x^{u^a}(\bar{k}) = 0_n$.

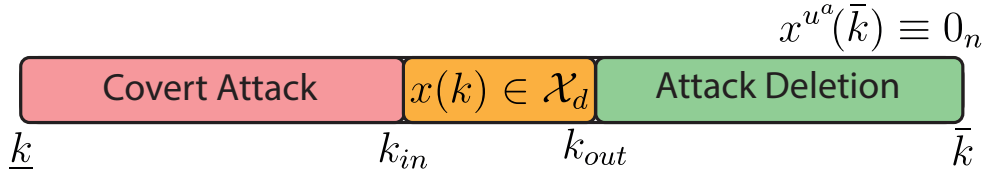


Figure 3.10: Finite-time attack: phases and actions.

3.3.6 Phase I - reaching $\mathcal{X}_d$

In this phase, the attack control input $u^a(k)$ is designed to replace $u(k)$ and robustly steer the plant's trajectory into the desired RCI region $\mathcal{X}_d$.

By resorting to the ST-MPC paradigm such a problem can be solved in finite-time as follows:

Offline attack preparation

By considering $\mathcal{X}_d$ as the terminal RCI region (target set) of the attacker, a family of robust one-step controllable sets $\{\mathcal{T}_d^i\}_{i=0}^{N_d}$, $N_d \geq 0$ is computed according to the following recursive definition:

$$\begin{aligned} \mathcal{T}_d^0 &:= \mathcal{X}_d \\ \mathcal{T}_d^i &:= \{x \in \mathbb{R}^n : \exists u^d \in \mathcal{U} \; s.t. \; Ax + Bu^d \in \tilde{\mathcal{T}}_d^{i-1}\}, \; i > 0 \end{aligned} \tag{3.30}$$

where $\tilde{\mathcal{T}}_d^i = \mathcal{T}_d^i \ominus B_d\mathcal{D}$, and $u_d \in \mathcal{U}$ is attack desired control input. Such a recursion is terminated when the union of controllable sets covers the admissible state space region $\mathcal{X}$, i.e.,

$$\mathcal{X} \subseteq \bigcup_{i=0}^{N_d} \mathcal{T}_d^i \tag{3.31}$$

Remark 3.7. *Please note that efficient tools and toolboxes exist to compute exact or approximated robust one-step controllable sets (3.30) for linear systems, see e.g., [75, 79, 88–91] and references therein.*

Attack actions ($\underline{k} \leq k \leq k_{out}$):

By taking advantage of the offline computations, the attacker's actions on the actuation channels reduce to the solution of a simple Quadratic Programming (QP) optimization problem forcing (at each step), the state trajectory to evolve within the family of controllable sets $\{\mathcal{T}_d^i\}_{i=0}^{N_d}$, until the terminal region $\mathcal{T}_d^0 \equiv \mathcal{X}_d$ is reached, i.e.,

$$\textit{if } x(k) \in \mathcal{T}_d^i \;\; \textit{compute } u^d(k) \in \mathcal{U} \textit{ s.t. } x(k+1) \in \mathcal{T}_d^{i-1} \tag{3.32}$$

As a consequence, the following FDI attack is performed to replace $u(k)$ with $u^d(k)$, i.e.,

$$u^a(k) = u^d(k) - u(k) \rightarrow u'(k) = u^d(k) \tag{3.33}$$

On the other hand, on the measurement channel, to avoid detection, the covert FDI in (3.29) is used.

The attacker's actions are summarized in the following algorithm:

Algorithm 3.1: Phase I (covert attack) - attacker's algorithm $\underline{k} < k \leq k_{out}$

Offline: Compute $\{\mathcal{T}_d^i\}_{i=0}^{N_d}$ as in (3.30)-(3.31)

Online: Compute $u^a(k), x^a(k)$ as follows:

1: Find the smallest set index $0 \leq i \leq N_d$ containing $x(k)$:

$$i(k) := \min_{0 \leq i \leq N_d} i : \; x(k) \in \mathcal{T}_d^i \tag{3.34}$$

2: **if** $i(k) == 0$ **then** $\mathcal{T}_d^{next} = \mathcal{T}_d^{i(k)}$

3: **else** $\mathcal{T}_d^{next} = \mathcal{T}_d^{i(k)-1}$

4: **end if**

5: Compute $u^d(k)$ solving the QP problem

$$u^d(k) = \arg\min_{u^d} \|Ax(k) + Bu^d - x_d\|_2^2 \quad s.t. \tag{3.35}$$

$$Ax(k) + Bu^d \in \tilde{\mathcal{T}}_d^{next},\ u^d \in \mathcal{U} \tag{3.36}$$

6: Determine $u^a(k)$ and $x^a(k)$ as in (3.33) and (3.29)

Lemma 3.1. *If the Phase I duration is greater or equal than N_d, i.e., $k_{out} - \underline{k} \geq N_d$, then Algorithm 3.1 ensures that the attack complies with the objective* ***(CH3.3-O1)*** *regardless of any admissible disturbance realization (3.18). Moreover, $k_{in} \leq \underline{k} + N_d$, and $x(k) \in \mathcal{X}_d, \forall\, k_{in} \leq k \leq k_{out}$.*

Proof - By construction, the QP optimization problem (3.35) is guaranteed to admit a solution $\forall\, k$ [79]. Moreover, if $x(k) \in \mathcal{T}_d^{i(k)}$ then $Ax(k) + Bu^d \in \tilde{\mathcal{T}}_d^{i(k)-1}$ and $x(k+1) \in \mathcal{T}_d^{i(k)-1}$. As a consequence, regardless of any initial condition $x(\underline{k}) \in \mathcal{X}$, the set-membership index $i(k)$ has a monotonically decreasing behavior until $i(k) = 0$ is reached. When $i(k) = 0$, then the attacker's control inputs aim to keep $x(k)$ into the RCI set $\mathcal{X}_d$. Therefore Algorithm 3.1 ensures that in the worst-case scenario $x(\underline{k} + N_d) \in \mathcal{T}_d^0 = \mathcal{X}_d$, $x(k) \in \mathcal{X}_d$, $\forall k_{in} \leq k \leq k_{out}$, where $k_{in} \leq \underline{k} + N_d$. □

Remark 3.8. *In the worst-case scenario, N_d time steps are needed to fulfill the requirements of Phase I (see Lemma 3.1). As a consequence, the duration of Phase I should be greater or equal to N_d.* □

3.3.7 Phase II - attack deletion ($x^{u^a}(\bar{k}) = 0_n$)

In phase II, the attacker, after achieving its primary objective (e.g., $x(k) \in \mathcal{X}_d$), wants to remove any trace of its presence to avoid detection in the post-attack phase. Specifically, as discussed in Remark 3.6, no passive anomaly detector (3.22) can discover anomalies in the post-attack phase if $x^{u^a} = 0_n, \forall\, k \geq \bar{k}$. Therefore, the attacker's action $u^a(k)$ for

$k_{out} < k \leq \bar{k}$ must be devoted to ensure that the state evolution due to the attacker actions (x^{u^a}) vanishes in a finite number of steps. Different from Phase I, where the attacker's actions aimed to replace $u(k)$ with $u^d(k)$ (see(3.33)), here the attacker wants to control only x^{u^a}. As a consequence, while removing x^{u^a} the attacker must make sure that the signal $u'(k) = u(k) + u^a(k)$ is admissible, i.e., $u'(k) \in \mathcal{U}$, regardless of the controller input $u(k)$ computed by (3.21).

Assumption 3.5. *There exists a small convex compact set $\Delta \subset \mathbb{R}^n$, $0_n \in \Delta$, s.t. such that*

$$u(k) \oplus \Delta \subseteq \mathcal{U}, \forall\, k \tag{3.37}$$

Remark 3.9. *Please note that such an assumption assumes that the control action $u(k)$ computed by (3.21) are contained into a proper inner set of $\mathcal{U}$. Such an assumption is reasonable in uncertain constrained setups where the controller actions are typically mapped into a smaller input set to ensure constraint satisfaction despite any disturbance realization (3.18) [92, 93]. Moreover, it is also fulfilled when the state trajectory is in proximity of the equilibrium state [92].* □

In what follows, Assumption 3.5 is instrumental to ensure that the attack deletion problem has a guaranteed solution in a finite number of steps. Then, in Remark 3.10, such an assumption is relaxed, and other conditions under which the attack deletion problem can be accomplished are investigated.

Offline attack deletion preparation

Lemma 3.2. *Consider the attacker's desired region $\mathcal{X}_d$, the Phase I attacker's algorithm, and $k_{out} \geq \underline{k} + N_d$. Then, regardless of any admissible disturbance (3.18) realization*

$$x^{u^a}(k_{out}) \in (\mathcal{X}_d \oplus -\mathcal{X}) := \mathcal{X}^{u^a} \tag{3.38}$$

Proof - According to (3.25), we have that

$$x^{u^a}(k_{out}) = x(k_{out}) - x^u(k_{out})$$

Moreover, by noticing that if $k_{out} \geq \underline{k} + N_d$ then $x(k_{out}) \in \mathcal{T}_d^0 \equiv \mathcal{X}_d$ and that $x^u(k_{out}) \in \mathcal{X}$, we have that (3.38) holds true, concluding the proof. □

By considering $x^{u^a}(\bar{k}) = 0_n$ as the target state and $\mathcal{X}^{u^a}$ as the initial admissible set for $x^{u^a}(k_{out})$, a family of robust one-step controllable sets in the attacker's state space x^{u^a}, namely $\{\mathcal{T}_a^j\}_{j=0}^{N_a}, N_a > 0$, is built considering $u^a(k) \in \Delta$ as the attacker's worst-case input constraint set, i.e.,

$$\begin{aligned} \mathcal{T}_a^0 :=& 0_n \\ \mathcal{T}_a^j :=& \{x^{u^a} \in \mathbb{R}^n : \exists u^a \in \Delta \; s.t. \; Ax^{u^a} + Bu^a \in \mathcal{T}_a^{j-1}\}, \; j > 0 \end{aligned} \tag{3.39}$$

Such a recursion is terminated when the admissible set of initial states $x^{u^a}(k_{out})$ is covered, i.e.,

$$\mathcal{X}^{u^a} \subseteq \bigcup_{i=0}^{N_a} \mathcal{T}_a^j \tag{3.40}$$

Lemma 3.3. *Under Assumption 3.5, if there exist a family of robust one-step controllable sets $\{\mathcal{T}_a^j\}_{j=0}^{N_a}$, built as in (3.39) and satisfying (3.40), then there exists a sequence of control inputs $\{u^a(k)\}_{k=k_{out}+1}^{\bar{k}-1}$ such that $x^{u^a}(\bar{k}) = 0_n$ and $u'(k) = u(k) + u^a(k) \in \mathcal{U}, \forall \, k_{out} < k \leq \bar{k} - 1$*

Proof - By construction, recursion (3.39) ensures that at each time steps there exists $u^a(k) \in \Delta$ such that the one-step evolution $x^{u^a}(k+1)$ belongs to a controllable set whose index is strictly lower than the current one, e.g., if $x^{u^a}(k) \in \mathcal{T}_a^j, \; j > 0 \rightarrow x^{u^a}(k+1) \in \mathcal{T}_a^{j-1}$. Therefore, recursively, we have that $x^{u^a}(\bar{k}) \in \mathcal{T}_a^0 = 0_n$. Moreover, according to Assumption 3.5, we are guaranteed that $u'(k) \in \mathcal{U}, \forall \, k_{out} < k \leq \bar{k} - 1$. □

Attack deletion ($k_{out} < k \leq \bar{k}$)

Similarly to what is done by the attacker in Phase I, in Phase II, the attacker's actions $u^a(k)$ and $x^a(k)$ are computed according to the following algorithm:

Algorithm 3.2: Phase II (attack deletion) - attacker's algorithm $k_{out} < k \leq \bar{k}$

Offline: Compute $\{\mathcal{T}_a^j\}_{k=0}^{N_a}$ as in (3.39)-(3.40)

Online: Compute $u^a(k), x^a(k)$ as follows:

1: Find the smallest set index $0 \leq j \leq N_d$ containing $x(k)$:

$$j(k) := \min_{0 \leq j \leq N_a} j : x(k) \in \mathcal{T}_a^j \tag{3.41}$$

2: **if** $j(k) == 0$ **then** $u^a(k) = 0_m$

3: **else**

4: Compute $u^a(k)$ solving the QP problem

$$\begin{aligned} u^a(k) &= \arg\min_{u^a} \|Ax^{u^a}(k) + Bu^a\|_2^2 \quad s.t. && (3.42)\\ & Ax^{u^a}(k) + Bu^a \in \mathcal{T}_a^{j(k)-1} && (3.43)\\ & u^a \in \mathcal{U} - u(k) && (3.44) \end{aligned}$$

5: **end if**

6: Determine $x^a(k)$ as in (3.29)

Lemma 3.4. *If the Phase II duration is greater than N_a, i.e., $\bar{k} - k_{out} > N_a$, then Algorithm 3.2 ensures that the attack is not detectable for $k \geq \bar{k}$, regardless of any admissible disturbance (3.18) realization.*

Proof - By following the same reasoning used in Lemma 3.1, if $\bar{k} - k_{out} > N_a$ then the monotonically decreasing set-membership index $j(k)$ is guaranteed to be zero for $k = \bar{k}$.

Therefore, since $\mathcal{T}_a^0 = 0_n$, we have that $\forall\, k \geq \bar{k}$, the contribution of the attack on the state of the system will be zero and detection in the post-attack phase is avoided. □

3.3.8 Proposed finite-time attack: feasibility and undetectability

In the following propositions, the properties of the finite-time attack developed in section 3.3.5 are investigated.

Proposition 3.2. *Consider the constrained plant model (3.17)-(3.20) and the anomaly detector (3.22). If, for a given target RCI set $\mathcal{X}_d$, there exist $0 \leq N_d < \infty$ such that (3.30) satisfies (3.31), $0 \leq N_a < \infty$ such that (3.39) complies with (3.40), and $\Delta \neq \emptyset$ in (3.37). Then, Algorithm 3.1 and Algorithm 3.2 ensure that:*

- *the finite-time covert attack (Phase I + Phase II) fulfills the objectives **(CH3.3-O1)**-**(CH3.3-O2)**, i.e., $\exists k : x(k) \in \mathcal{X}_d$ and the attack is undetectable by (3.22) for $k > \underline{k}$.*
- *irrespective of any admissible initial plant condition $x(\underline{k}) \in \mathcal{X}$ and bounded disturbance realization $d(k) \in \mathcal{D}$, the minimum attack duration $\bar{T}$ to fulfill **(CH3.3-O1)**-**(CH3.3-O2)** is $\bar{T} = N_d + N_a$.*

Proof - By collecting the results in Lemmas 3.1-3.4, Algorithm 3.1 ensures undetectability for $\underline{k} \leq k \leq k_{out}$ and that $x(k) \in \mathcal{X}_d$ for $k_{in} \leq k \leq k_{out}$. Moreover, Algorithm 3.2 guarantees that the post-attack undetectability condition $x^{u^a}(\bar{k}) = 0_n$ is reached for $k \geq k_{out} + N_a$. Therefore, the minimum finite-time attack duration that ensures fulfilling **(CH3.3-O1)**-**(CH3.3-O2)** regardless of $x(\underline{k}) \in \mathcal{X}$ is obtained for $k_{out} = k_{in}$ and $\bar{T} = N_d + N_a$. □

Proposition 3.2 implies that if the attack duration, namely $\bar{T}$, is bigger or equal to $N_a + N_d$, then the proposed finite-time covert-attack is feasible starting from any $x(k) \in \mathcal{X}$.

In the next proposition, this is formalized, and it is also shown that for $\bar{T} < N_a + N_d$, the attack might still be feasible starting from a subset of $\mathcal{X}$.

Proposition 3.3. *Given a desired attack duration $\bar{T}$, the set of initial state condition $\mathcal{X}_a \subseteq \mathcal{X}$, $x(\underline{k}) \in \mathcal{X}_a$ such that finite-time attack (Algorithms 3.1-3.2) can be successfully completed in $\bar{T}$-steps can be offline determined and it is equal to:*

$$\mathcal{X}_a = \left(\bigcup_{i=0}^{\min(\bar{T}-N_a, N_d)} \mathcal{T}_d^i \right) \bigcap \mathcal{X} \tag{3.45}$$

Proof - First, it is important to underline that regardless of the initial state condition, the attack duration cannot be lower than N_a (number of steps required to cancel out the presence of the attack in Phase II). Therefore, the number of steps available to the attacker to steer $x(\underline{k})$ into $\mathcal{X}_d$ is $\bar{T} - N_a$. As a consequence, since $\mathcal{X} \subseteq \bigcup_{i=0}^{N_d} \mathcal{T}_d^i$, if $\bar{T} \geq N_d + N_a$, then $\min(\bar{T} - N_a, N_d) = N_d$ and the set of admissible initial condition is equal to entire set of admissible states, i.e., $\mathcal{X}_a = \mathcal{X}$. On the other hand, if $\bar{T} < N_a + N_d$, then $\min(\bar{T} - N_a, N_d) = \bar{T} - N_a$ and $\mathcal{X}_a \subset \mathcal{X}$, see Figure. 3.11 for an illustration. □

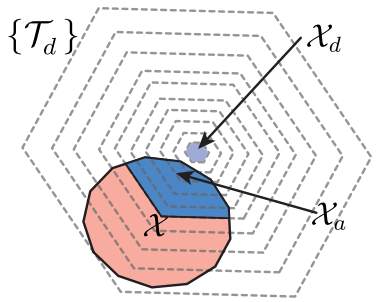


Figure 3.11: The state subspace $\mathcal{X}_a \subseteq \mathcal{X}$ (blue region) from which the attack can successfully perform the finite-time attack for $\bar{T} < N_d + N_a$.

Remark 3.10. *The finite-time attack is guaranteed to exist under Assumption 3.5, i.e., $\Delta \neq \emptyset$, $\forall\, k_{out} < k \leq \bar{k}$. However, it is important to underline that this is only a sufficient condition and that the attack might be feasible otherwise. For the sake of completeness and to open the floor to further research directions, three different situations can be analyzed:*

- *Consider the case where $\Delta = \emptyset$ and A is Nilpotent with index N, i.e., $A^N = 0_{n\times n}$. In this case, since $x^{u^a}(k) = \sum_{j=0}^{k-1} A^j B^j u^a(k-1-j)$, then, regardless of the initial attack state $x^{u^a}(k_{out})$, $x^{u^a}(k_{out}+N) = 0_n$ if $u^a(k) = 0_m, \forall\, k \geq k_{out}$. Therefore, in Phase II, the attacker does not need to take any actions on the actuation channel to ensure that $x^{u^a}(k)$ converges to zero in N-steps (see (3.27)). In particular for $k_{out} < k \leq \bar{k}$, $\bar{k} - k_{out} > N$, the attacker can use Algorithm 3.2 where in Step 4 the optimization problem (3.42)-(3.44) is replaced by $u^a(k) = 0_m$.*

- *Consider the case where $\Delta = \emptyset$ and the matrix A is Schur stable, i.e., its eigenvalues have modulus less than 1. In this case, in Step 4 of Algorithm 3.2, the attacker can evaluate if the optimization problem (3.42)-(3.44) admits a solution for the input constraint $u^a(k) \in \mathcal{U} - u(k)$. If such a problem does not admit a solution then the attacker can apply $u^a(k) = 0_m$, and exploit the contracting nature of A. In such a circumstances, the attack is guaranteed to end when the optimization problem (3.42) admits a solution for at-most N_a time steps. However, in this case it is not possible to offline determine the number of steps needed to complete Phase II.*

- *Consider the case where $\Delta = \emptyset$ and the matrix A is unstable. In this case, it is not possible to guarantee that the attack can terminate in a finite-amount of time. Furthermore, x^{u^a} is not guaranteed to remain inside $\bigcup_{j=0}^{N_a}\{\mathcal{T}_a^j\}$ and the recursive feasibility of Algorithm 3.2 is not ensured.*

Remark 3.11. *The proposed finite-time attack has been designed under the assumption that the entire state vector can be measured. Nevertheless, such an attack can be also designed to deal with a plant model* (3.17) *characterized by an output equation $y(k) = Cx(k) + d_y(k)$, where $C \in \mathbb{R}^{p\times n}$, $y(k) \in \mathbb{R}^p$ is the sensor measurement vector, and $d_y(k) \in \mathcal{D}_y \subset \mathbb{R}^p$ is a compact but unknown measurement disturbance set containing the origin. In general, if $C \neq I$, the extension is possible if (i) a state-estimator capable of dealing with bounded process and measurement disturbances can be designed, (ii) the worst-case state-estimation error can be characterized. The first is needed to reconstruct $x(k)$, while*

the second is important to properly build a family of robust one-step controllable sets (see e.g., (3.30)) that takes into account the bounded errors introduced by the estimator. Please refer to, e.g., [75, Chapter 11] and reference therein, for exhaustive details on the design of state estimators fulfilling the requirements (i)-(ii). On the other hand, a straightforward extension can be provided if $C = I$ (i.e., the entire state vector can be measured with a bounded error). Note that in this particular case, if $y(k)$ is measured, then $x(k)$ is also known with some uncertainty, i.e., $x(k) \in y(k) \oplus (-\mathcal{D}_y)$. Therefore, such extra uncertainty can be then taken into account in the construction of the robust one-step controllable sets (3.30) by simply computing $\tilde{\mathcal{T}}_d = \mathcal{T}_d \ominus (B_d\mathcal{D} \oplus (-A\mathcal{D}_y))$, see [94]. □

3.3.9 Simulation Example

In this section, the industrial Continuous-Stirred Tank Reactor (CSTR) system used in [95, 96] and shown in Figure. 3.12 has been considered to show in simulation the effectiveness of the proposed constrained finite-time attack.

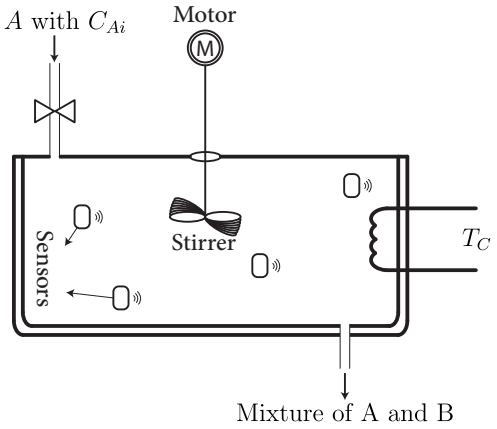


Figure 3.12: Continuous-Stirred Tank Reactor (CSTR) system

In this system, the chemical species $\mathcal{A}$ react with the chemical species $\mathcal{B}$ at a specific temperature. The output of the system is a mixture of these two chemicals (see Figure. 3.12). The state vector of the CSTR system is $x = [C_{\mathcal{A}}, T_r]^T$ where $C_{\mathcal{A}}$ is the concentration of the chemical species $\mathcal{A}$, and T_r is the reaction temperature. On the other hand, $u = [T_C, C_{\mathcal{A}i}]^T$ is the control vector where T_C is the cooling controlled temperature

and $C_{\mathcal{A}i}$ is the input concentration of the chemical species $\mathcal{A}$. The dynamics of the CSTR system has been linearized and discretized with a sampling time $T_s = 1s$ and the resulting A, B and B_d matrices are:

$$A = \begin{bmatrix} 0.9719 & -0.0013 \\ -0.0340 & 0.8628 \end{bmatrix}, B = \begin{bmatrix} -0.0839 & 0.0232 \\ 0.0761 & 0.4144 \end{bmatrix} \\ B_d = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{3.46}$$

The admissible disturbance set is

$$\mathcal{D} = \{d \in \mathbb{R}^2 : \begin{bmatrix} -0.01 \\ -0.08 \end{bmatrix} \leq d \leq \begin{bmatrix} 0.01 \\ 0.08 \end{bmatrix}\} \tag{3.47}$$

and the states and inputs are subject to the following constraint sets:

$$\begin{aligned} \mathcal{X} &= \{[C_{\mathcal{A}}, T_r]^T \in \mathbb{R}^2 : -2 \leq C_{\mathcal{A}} \leq 2, -10 \leq T_r \leq 10\} \\ \mathcal{U} &= \{[T_C, C_{\mathcal{A}i}]^T \in \mathbb{R}^2 : -2 \leq T_C \leq 2, -2 \leq C_{\mathcal{A}i} \leq 2\} \end{aligned} \tag{3.48}$$

The controller (3.21) is a stabilizing state-feedback controller $u(k) = -K(x - r(k)) + u_{eq}(k)$, with u_{eq} the equilibrium input associated to the desired equilibrium state $x_{eq} = r(k)$, and K (the controller gain) is:

$$K = \begin{bmatrix} -10.903 & 0.560 \\ 1.921 & 1.978 \end{bmatrix} \tag{3.49}$$

Moreover, the used set Δ is:

$$\Delta = \{[T_C, C_{\mathcal{A}i}]^T \in \mathbb{R}^2 : -0.5 \leq T_C \leq 0.5, -0.5 \leq C_{\mathcal{A}i} \leq 0.5\} \tag{3.50}$$

The finite-time attack is offline configured as follows. The attacker wants to steer the state of the system in the proximity of the equilibrium pair (x_d, u_d), where $x_d = [-2.5, 0]^T$

and $u_d = [0.74, -0.34]^T$ Please note that the equilibrium state is outside of the admissible safe region $\mathcal{X}$. Moreover, the desired RCI region $\mathcal{X}_d$, centered in x_d, is computed as in [97] (see the blue region in Figure. 3.14). Then, the attacker builds the families of robust one-step controllable sets $\{\mathcal{T}_d^i\}_{i=0}^{N_d}$ (see Figure. 3.14) and $\{\mathcal{T}_a^j\}_{j=0}^{N_a}$ (see Figure. 3.15) as in (3.30) and (3.39), respectively. In particular, the terminal conditions (3.31) and (3.40) are reached for $N_d = 28$ and $N_a = 47$. As a consequence, the minimum number of steps required to complete the attack for any $x(\underline{k}) \in \mathcal{X}$ is $N_d + N_a = 75$ (see Proposition 3.2 and Figure. 3.15). Please note that by exploiting the result in Proposition 3.3, given a finite duration $\bar{T}$, the attacker is able to offline determine the sets of states $\mathcal{X}_a \subseteq \mathcal{X}$ from where the attack can be successfully completed. In Figure. 3.15, $\mathcal{X}_a$ is shown for $\bar{T}$ equals to 60, 70 and 75.

In the carried out simulation, the attacker launches for two times the finite-time covert attack described by Algorithms 3.1 and 3.2. The details of the attacks, i.e., $\underline{k}$, $k_{in}, k_{out}, \bar{k}$ are shown in Table 3.1. The plant initial condition is $x(0) = 0_2$, and $r(k)$ is shown in Figure. 3.13.

Table 3.1: Finite-time covert attacks timing information

	first attack	second attack
$\underline{k}$	31 s	200 s
k_{in}	53 s	223 s
k_{out}	59 s	234 s
$\bar{k}$	72 s	245 s

Figure. 3.13 shows the evolution over time for the two components of $r(k)$, $x(k)$ and $x'(k)$. It is possible to notice that during the two attacks $x(k)$ deviates significantly from $r(k)$ causing a constraint violation for $31 \leq k \leq 72$ and $200 \leq k \leq 245$. On the other hand $x'(k)$ (i.e., the signal received by the controller and detector) is unaffected by the presence of the attack. Moreover, the difference between $x'(k)$, and $x(k)$, i.e., the attack's state $x^{u^a}(k)$ becomes exactly zero when each attacks are terminated at $k = 72$ and $k = 245$,

respectively. As a consequence, the designed attack can be repeatedly executed avoiding detection during the attack and afterwards.

To better appreciate the *modus operandi* of the attack, Figure. 3.14 and Figure. 3.15 show the state trajectory of the plant ($x(k)$) and attacker ($x^{u^a}(k)$). In Figure. 3.14, the state trajectory has been divided in four different colors to better highlight the phases of the two attacks. Regardless of the state $x(\underline{k})$, it is possible to notice that the attacker is capable of steering the trajectory in the RCI region $\mathcal{X}_d$. In particular, as shown in Table 3.1, for the first and second attack, the RCI region is reached in 12 and 23 steps, respectively. Moreover, as better shown in Figure. 3.15, in Phase II, regardless $x^{u^a}(k_{out}) \in \mathcal{X}_d \oplus (-\mathcal{X})$, the attack termination condition $x^{u^a} = 0_2$ is reached in a finite number of steps. Moreover, the time required for the attacker to completely execute attack 1 and 2 (i.e., $\bar{k} - \underline{k}$) is equal to $41s$ and $45s$, respectively. Such a duration is lower of the worst-case execution time of $75s$ that can be offline predicted by the attacker using (3.45), see e.g., $\mathcal{X}_a$ in Figure. 3.14. Finally, in Figure. 3.16, the control signal $u(k)$, $u'(k)$ and $u^a(k)$ are shown. It is possible to appreciate that the attacker's input vector $u^a(k)$ always ensures that the control signal received by the plant, i.e $u'(k)$, complies with the input constraints.

3.4 Conclusion

In this chapter, we have shown the existence of finite-time attacks against constrained and unconstrained CPSs. The proposed finite-time 3-phase covert attack against unconstrained control systems first exploits an FIR state estimator to obtain, in finite-time, unbiased estimation of the system's states. Then, it uses reachability arguments to design attack vectors capable of arbitrarily alter the plant state trajectory and ensuring undetectability once the attack is ended. On the other hand, the proposed fine-time covert attack against constrained control systems has been designed by jointly combining robust controllability arguments and a set-theoretic-based receding horizon control paradigm. It has been formally proved that both of the designed attacks are stealthy regardless of any

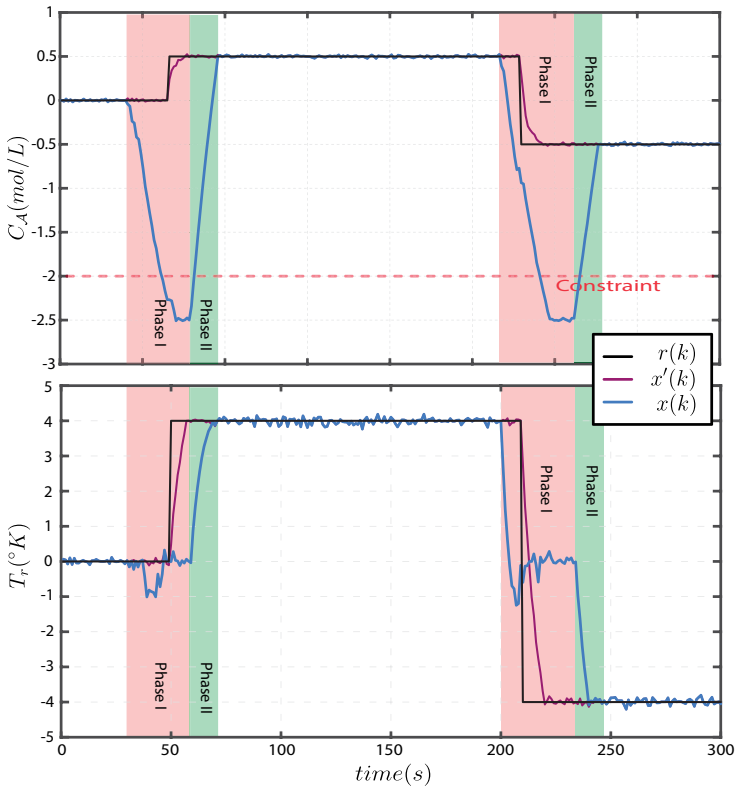


Figure 3.13: CSTR Plant's states evolution in the presence of the finite-time covert attack

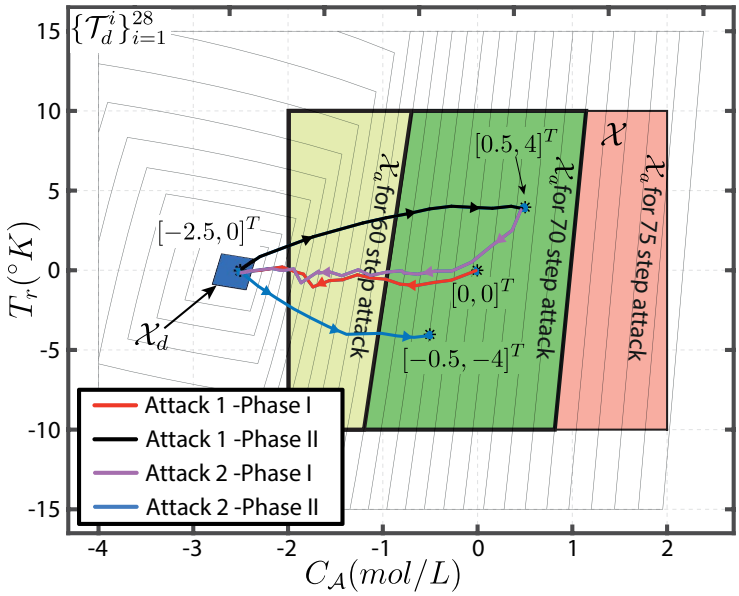


Figure 3.14: Plant's state trajectory $x(k)$ and family of robust one-step controllable sets $\{\mathcal{T}_d^i\}_{i=0}^{25}$. The yellow, green and pink regions inside $\mathcal{X}$ depict the set of initial states $\mathcal{X}_a \subseteq \mathcal{X}$ for which the constrained finite-time attack can be completed if $\bar{T} = 60$ (yellow region), $\bar{T} = 70$ (yellow + green regions) and $\bar{T} = 75$ (yellow + green + pink regions).

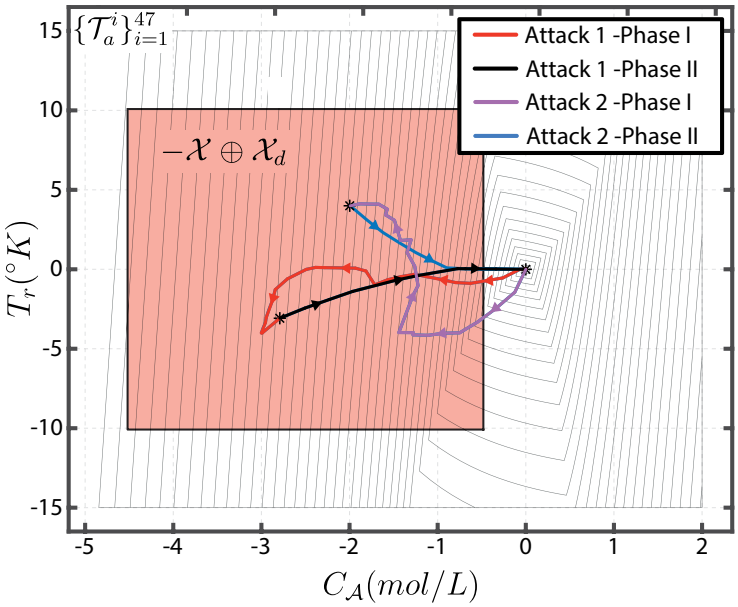


Figure 3.15: Attacker's state trajectory (x^{u^a}) and family of robust one-step controllable sets $\{\mathcal{T}_a^i\}_{i=0}^{47}$.

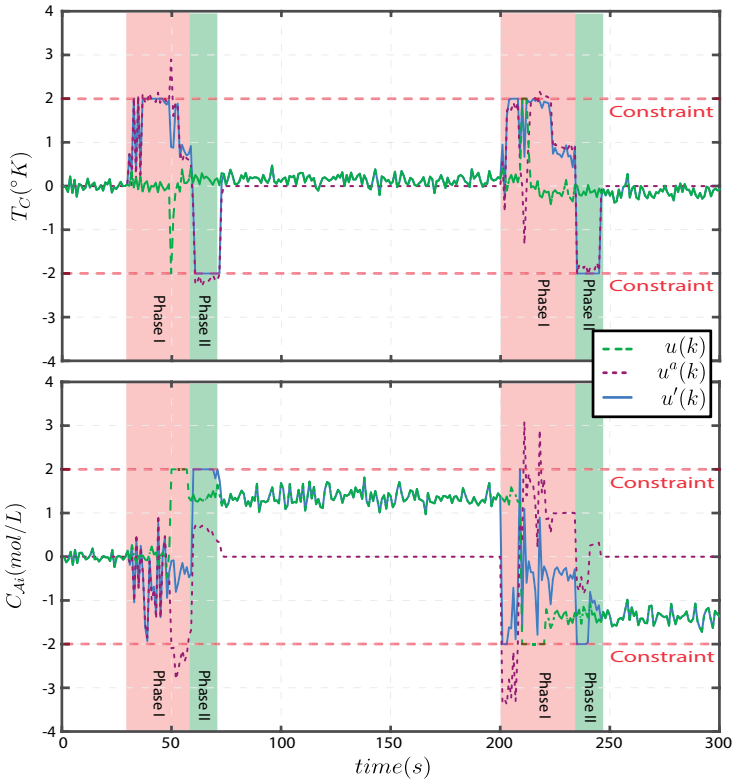


Figure 3.16: Control signals $u(k)$, $u^a(k)$ and $u'(k)$.

anomaly detector deployed on the controller side of the networked CPS.

Chapter 4

A Safety Preserving Control Architecture for Cyber-Physical Systems

4.1 Introduction

In this chapter, we propose a networked control architecture to ensure the plant's safety in the presence of cyber-attacks on the communication channels. In particular, we consider systems subject to both state and input constraints that must be preserved for safety reasons despite any admissible attack scenario. To this end, first, two different detectors are proposed to detect attacks on the setpoint signal as well as on the control inputs and sensor measurements. Then, an emergency controller, local to the plant, is designed to replace the networked controller whenever an attack is detected. Finally, the concept of robust N-step attack-safe region is introduced to ensure that the emergency controller is activated, regardless of the detector performance, at least one-step before the safety constraints are violated. It is formally proved that the plant trajectory is uniformly ultimately bounded in an admissible region regardless of the attacker's actions and duration. Finally, by considering a continuous-stirred tank reactor system, numerical simulations

are presented to show the proposed solution's capabilities.

4.1.1 Contribution of the work

By referring to the networked control scheme in Figure. 4.1, three main contributions of this research study are the followings. i) The proposed solution is capable of detecting FDI attacks on the reference (setpoint) signal. To the best of our knowledge, this problem has not received sufficient attention in the literature. The main difficulty to detect such attacks is given by the absence of a-priori information on its expected time evolution which typically is not a function of the dynamics of the closed-loop system. Setpoint detection is here achieved by proposing a different networked architecture where the reference signal becomes a function of the state of the system. ii) The proposed solution ensures that plant's safety and uniform ultimate boundedness of the state trajectory are preserved, regardless of the attacker's actions and irrespective of the detector's performance. Differently from the competitor solutions in [59, 98], the attacker's actions are not limited to be a function of the state of the system. Contrary to [61], the proposed solution does not require that attack-free communication channels to be re-established in a-priori known number of steps. iii) The concept of $N-$step attack safe region is for the first time introduced to design an emergency controller capable of replacing the networked controller in the presence of unreliable communication channels.

4.2 Networked Control System Setup

Consider the following class of discrete-time linear time-invariant (LTI) systems

$$x_p(k+1) \;=\; Ax_p(k) + Bu(k) + B_d d(k) \tag{4.1}$$

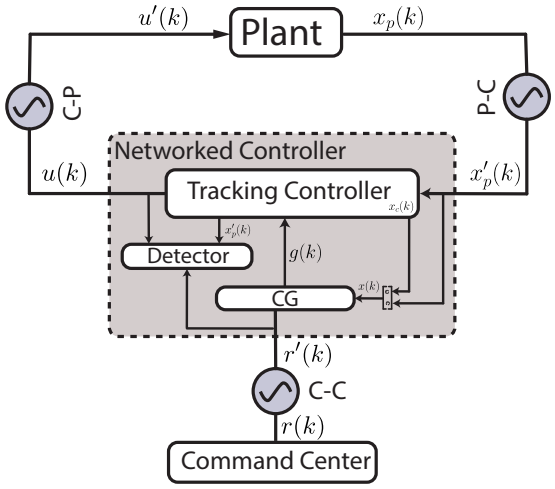


Figure 4.1: Considered Networked Control System Setup

where $x_p(k) \in \mathbb{R}^{n_p}$ is the state vector, $u(k) \in \mathbb{R}^m$ is the input vector and $d(k)$ is a bounded disturbance:

$$d(k) \in \mathcal{D} \subset \mathbb{R}^d, \quad 0_d \in \mathcal{D} \tag{4.2}$$

Due to actuation limitations the input signal is constrained into the following set

$$u(k) \in \mathcal{U}, \quad \forall\, k \in \mathbb{Z}_+ \tag{4.3}$$

where $\mathcal{U} \subseteq \mathbb{R}^m$ is a convex compact subset with $0_m \in \mathcal{U}$. Moreover, (4.1) is subject to the following state constraint

$$x_p(k) \in \mathcal{X}_p \quad \forall\, k \in \mathbb{Z}_+ \tag{4.4}$$

with $\mathcal{X}_p \subseteq \mathbb{R}^{n_p}$ a convex compact subset with $0_{n_p} \in \mathcal{X}$.

4.3 Command Governor (CG) Tracking Controller

In this section, the operations of the Command Governor (CG) tracking controller [46–48] (see the subsystems Tracking Controller and CG in Figure. 4.1) are revised. The control scheme consists of two nested loops:

- Inner Loop (primal controller): a stabilizing offset-free linear reference $r(k) \in \mathbb{R}^p$ tracking controller for the unconstrained and disturbance-free plant model (4.1).

- Outer Loop (CG): a supervisory controller that, whenever necessary, is in charge of modifying the received reference signal $r(k)$ to avoid constraints violation regardless of any admissible disturbance realization.

Under the actions of the primal controller and CG, the closed-loop plant (4.1) dynamics can be described as follows

$$\begin{aligned} x(k+1) &= \Phi x(k) + Gg(k) + G_d d(k) \\ c(k) &= H_c x(k) + Lg(k) + L_d d(k) \end{aligned} \tag{4.5}$$

where $x(k) = [x_p^T(k),\, x_c^T(k)]^T \in \mathbb{R}^n$, $n = n_p + n_c$, is the closed-loop state vector, $x_c(k) \in \mathbb{R}^{n_c}$ the state of the primal controller, and Φ, G and G_d are the closed-loop system matrices. Moreover $g(k) \in \mathbb{R}^{n_p}$ is the CG output signal and $c(k) \in \mathbb{R}^{n_c}$ the constrained vector embedding be prescribed state and input constraints (4.3)-(4.4)

$$c(k) \in \mathcal{C} \;\; \Leftrightarrow \;\; (u(k) \in \mathcal{U} \wedge x_p(k) \in \mathcal{X}_p),\; \forall k \geq 0, \tag{4.6}$$

with $\mathcal{C}$ a convex and compact set and H_c, L and L_d matrices of suitable dimensions.

The set of all constant virtual commands whose state evolution starting from x satisfies all the constraints is given by

$$\mathcal{V}(x) = \left\{\omega \in \mathcal{W}^\delta : \bar{c}(k, x, \omega) \in \mathcal{C}_k, \forall k = 0, \ldots, k_0\right\} \tag{4.7}$$

where

$$\mathcal{W}^\delta := \left\{\omega \in \mathbb{R}^m : \bar{c}_\omega \in \mathcal{C}^\delta\right\} \supset \mathcal{V}_i(x), \text{ with } \; \mathcal{C}^\delta = \mathcal{C}_\infty \ominus \mathcal{B}^\delta$$

is the set of all commands whose corresponding steady-state solutions satisfy the constraints with a tolerance margin δ (i.e., $\mathcal{C}^\delta \oplus \mathcal{B}^\delta \subset \mathcal{C}_\infty$), and k_0 is the prediction horizon

computed as in [48]. Furthermore, $\mathcal{B}^{\delta}$ is a ball of radius δ centered at the origin, and the sets $\mathcal{C}_k$ and $\mathcal{C}_\infty$ are obtained by means of the following Minkowski difference recursions

$$\begin{gathered}\mathcal{C}_0 := \mathcal{C} \ominus L_d\mathcal{D}, \quad \mathcal{C}_k := \mathcal{C}_{k-1} \ominus H_c\Phi^{k-1}G_d\mathcal{D} \\ \mathcal{C}_\infty := \bigcap_{k=0}^{\infty} \mathcal{C}_k\end{gathered} \tag{4.8}$$

and

$$\begin{gathered}\bar{c}(k,x,\omega) = H_c\left(\Phi^k x + \sum_{r=0}^{k-1}\Phi^{k-r-1}G\omega\right) + L\omega, \\ \bar{c}^{\omega} := H_c(I_n - \Phi)^{-1}G\omega + L\omega\end{gathered} \tag{4.9}$$

At each time instant k, the CG computes the best approximation $g(k)$ of the setpoint signal $r'(k)$ solving the following quadratic programming problem

$$g(k) = \arg\min_{\omega\in\mathcal{V}(x(k))} ||\omega - r'(k)||^2 \tag{4.10}$$

Remark 4.1. *For the aim of this chapter, it is particularly important to underline the following properties of the CG control scheme*

1. *By defining the set of admissible states and references as*

$$\mathcal{Z} := \{[r^T, x^T]^T \in \mathbb{R}^r \times \mathbb{R}^n \,|\bar{c}(k,x,r) \in \mathcal{C}^{\delta}, \forall k \in \mathbb{Z}_+\} \tag{4.11}$$

then, the projection of $\mathcal{Z}$ onto x_p defines the controller's Domain of Attraction (DoA), namely $\mathcal{X}_c \subseteq \mathcal{X}_p$, where the controller is capable of tracking the reference signal and fulfilling the prescribed constraints

$$\mathcal{X}_c := \{x_p \in \mathbb{R}^{n_p} \;\; s.t. \;\exists r \in \mathbb{R}^r,\, x_c \in \mathbb{R}^{n_c} : [r^T, [x_p^T, x_c^T]]^T \in \mathcal{Z}\} \tag{4.12}$$

2. *At each time k, the optimization (4.10) provides the best feasible approximation $g(k)$ of $r'(k)$.*

3. *If* $r'(k) = \bar{r}'_i \in \mathbb{R}^r$ *for* $i\gamma \leq k < (i+1)\gamma$, $\gamma > 0$, $i \geq 0$ *then* $||g(k)||$ *is monotonically non-decreasing (MN-D) or monotonically non-increasing (MN-I) for* $i\gamma \leq k < (i+1)\gamma$.

Please refer to [47, 48] for a detailed discussions about all the properties of the CG control paradigm. □

Definition 4.1. *Given the networked control scheme in Figure. 4.1 and the tracking controller domain* $\mathcal{X}_c \subseteq \mathcal{X}_p$, *the system (4.1) is called safe in the presence of any cyber-attack in the communication channels if*

$$x_p(k) \in \mathcal{X}_c, \ \forall\, k. \tag{4.13}$$

4.4 Problem Formulation

Consider the networked control architecture shown in Figure. 4.1 and the following FDI attacks on the control input ($u(k)$), state ($x_p(k)$) and setpoint ($r(k)$) signals

$$\begin{aligned} u'(k) &= u(k) + u^a(k) \\ x'_p(k) &= x_p(k) + x^a_p(k) \\ r'(k) &= r(k) + r^a(k) \end{aligned} \tag{4.14}$$

where $u^a(k) \in \mathbb{R}^m$, $x^a_p(k) \in \mathbb{R}^{n_p}$, and $r^a(k) \in \mathbb{R}^r$ are arbitrarily possibily unbounded vectors. The system (4.1) evolution, in the presence of FDI attacks, is

$$x_p(k+1) = Ax_p(k) + Bu'(k) + B_d d(k) \tag{4.15}$$

Assumption 4.1. *The attacker aims to inject the vectors* $u^a(k)$, $x^a_p(k), r^a(k)$ *to steer the system trajectory outside of the controller's DoA (4.12) and violate the safety constraints (4.3)-(4.4) while remaining undetectable.* □

Assumption 4.2. *The setpoint signal is assumed to be piecewise constant, i.e.,*

$$r(k) = \bar{r}_i, \quad i\gamma \leq k < (i+1)\gamma, \quad \forall\, i \in \mathbb{Z}_+$$

with $\bar{r}_i \in \mathbb{R}^r$ *and* $\gamma > 0$ *an a-priori known time interval* □.

Remark 4.2. *Please note that Assumption 4.2 is without loss of generality. Indeed, if* $\gamma = 1$, *then the reference signal can be changed at each sampling time.*

Assumption 4.3. *In the presence of unreliable network communications, we assume that at the cyber-layer, the networked controller can interrupt the data communications [99].* □

The control problem can be stated as follows:
Consider the networked control system in Figure. 4.1, the networked tracking controller's DoA (4.12), the FDI attack model (4.14), and the plant model (4.15). Design a networked control architecture capable of

- ***(CH4-O1)*** *detecting FDI attacks (4.14) before an unsafe configuration, violating the safety constraint (4.13), is reached;*
- ***(CH4-O2)*** *ensuring that the constraints (4.3)-(4.4) are always fulfilled and that normal operations can be restored once the attack-free scenario is recovered;*

4.5 Proposed Networked Control Architecture

First, to motivate the proposed control architecture (Figure. 4.2), it is important to underline the limitations of networked scheme in Figure. 4.1 to solve the objectives **(CH4-O1)**-**(CH4-O2)** :

1. The setpoint signal $r(k)$ evolution is not a function of the closed-loop control system operations. As a consequence, it is not possible to exploit control-theoretical tools

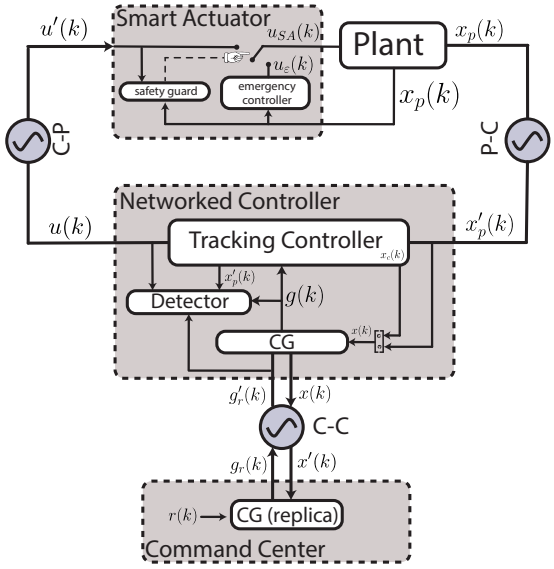


Figure 4.2: Proposed Control Architecture

to detect the presence of setpoint attack in the C-C channel, see [45] for a detailed discussion.

2. Detectors, whose actions are only performed in the control center, cannot detect advanced coordinated stealthy attacks in the C-P and P-C channels, see, e.g., the covert attack described in [26]. Consequently, some attacks cannot be detected even if the plant's state is driven by the attacker, outside of the networked controller domain $\mathcal{X}_c$.

3. If an attack is affecting the actuation channel (C-P), no emergency control actions can be taken in the Networked Controller (e.g., the attacker can intercept $u(k)$ and replace it with an arbitrary malicious input).

To mitigate the above drawbacks, the control architecture depicted in Figure. 4.2 is proposed. In particular,

- A CG replica is added to the command center to enable the detection of setpoint attacks.

- A Smart Actuator module is added locally to the plant to:

– Perform safety checks to guarantee that attacks are detected, in the worst-case scenario, one-step before the safety of the system is compromised (safety guard);

– Replace the networked controller with a local emergency controller whenever an attack is sensed. The emergency controller is designed to steer the system's state within the safe region $\mathcal{X}_e$ while fulfilling the prescribed safety constraints (4.3)-(4.4).

The next subsections are devoted characterizing the *modus operandi* of the proposed architecture.

4.5.1 Detector

The proposed detector exploits two different anomaly detection rules to detect FDI attacks on the C-P, P-C and C-C channels. In particular, detection of attacks on the control signal and state measurements is achieved using a robust anomaly detector based on reachability arguments, while attacks on the setpoint signal will be detected by exploiting basic properties of the CG control paradigm (see Remark 4.1).

4.5.1.1 Detection of FDI on the C-P and P-C channels

To detect attacks on the C-P and P-C channels, we define the expected system's robust one-step evolution as follows

$$\mathcal{X}^+(x'_p, u) = \{x^+ \in \mathcal{X}_p \text{ s.t. } x_p^+ = Ax'_p + Bu + B_d d, \forall d \in \mathcal{D}\} \tag{4.16}$$

Given the expected robust evolution at the time step $k-1$, i.e., $\mathcal{X}^+(x'_p(k-1), u(k-1))$, the first anomaly detection rule is the following

$$\textit{Rule-1}\,(k) = \begin{cases} \text{anomaly}, & \textit{if } x'_p(k) \notin \mathcal{X}^+(x'_p(k-1), u(k-1)) \\ \text{normal}, & \textit{otherwise} \end{cases} \tag{4.17}$$

Remark 4.3. *In the literature, different detectors have been proposed to reveal attacks on the measurement and actuation channels [34], and most of the existing solutions can be used in the proposed architecture. However, given the uncertain and constrained system model (4.1), the proposed robust detection rule (4.17) has the advantage to be, by design, robust against false positives. This feature arises from the construction of the system's robust one-step evolution set* $\mathcal{X}^+(x'_p, u)$ *exploited by (4.17), i.e., an anomaly is triggered if* $x'_p(k) \notin \mathcal{X}^+(x'_p(k-1), u(k-1))$. *Since* $\mathcal{X}^+(x'_p, u)$ *is built taking into account the worst-case disturbance realization* $\mathcal{D}$*, then it is not possible, in an attack-free scenario, that the one-step system evolution will be outside* $\mathcal{X}^+(x'_p, u)$. *In the proposed solution, the absence of false positives is desired to avoid triggering network disconnections and the emergency controller due to the presence of plant's disturbances (4.2).* □

4.5.1.2 Detection of FDI on the C-C channel

If a replica of the CG optimization (4.10) is performed in the control center (see Figure. 4.2), then, as shown in [45], the setpoint signal received by the CG and Detector, namely $g'_r(k)$, becomes a function of the closed-loop system operations and FDI attacks detection can be achieved. In particular, by observing the optimization problems solved by the CG and CG replica

$$\text{(CG replica):} \quad g_r(k) = \arg \min_{\omega \in \mathcal{V}(x'(k))} ||\omega - r(k)||^2 \tag{4.18}$$

$$\text{(CG):} \quad g(k) = \arg \min_{\omega \in \mathcal{V}(x(k))} ||\omega - g_r(k)||^2 \tag{4.19}$$

it is possible to appreciate that in the absence of FDI attacks, the CG optimization resolve the same optimization problem solved by the CG-replica. In particular, it solves an optimization problem starting from the solution given by the CG-replica, i.e., $g_r(k)$. Therefore, in the absence of attacks, despite any disturbance realization, $g(k) \equiv g_r(k), \forall\, k$ [45].

Proposition 4.1. *Let us assume that the CG and CG replica are identically designed according to (4.5)-(4.10). Then, in the absence of attacks, the followings hold true:*

(a) *The signal* $g(k) \equiv g_r(k)$, $\forall k$;

(b) *Given a fixed reference signal* $r(k) \equiv r$*, then* $g(k)$ *and* $g_r(k)$ *have a MN-I or MN-D behavior.*

Proof. The proposition can be demonstrated by resorting to a proof by contradiction (*Reductio ad Absurdum*).

(**a**) - Let us assume that at the generic time instant k, the setpoint $r(k)$ is imposed and $g(k)$ and $g_r(k)$ are such that $g(k) \neq g_r(k)$. By virtue of the CG properties in the remark 4.1, $g_r(k)$ uniquely exists and it represents the best feasible approximation of the setpoint $r(k)$. By following similar reasonings, $g(k)$ is the best feasible approximation of $g_r(k)$. Therefore, if $g_r(k) \neq g(k)$ (hypothesis) we can imply that $g_r(k)$ is not the best feasible approximation of $r(k)$ and we reach a contradiction (absurd). As a consequence, the only possibility is that $g(k) \equiv g_r(k)$ $\forall k$.

(**b**) - Let us consider a fixed setpoint $r(k) \equiv \bar{r}$, $\forall k \geq \bar{k}$ with $\bar{r} \in \mathcal{W}^{\delta}$ (see (4.7)) and $g(k), g_r(k)$, are such that $g_r(\bar{k}) < \bar{r}$, and $g(\bar{k}) < \bar{r}$. Moreover, let us pick two generic time instants k_1 and k_2 such that $\bar{k} \leq k_1 < k_2$ and make the hypothesis $g_r(k_2) < g_r(k_1)$, and $g(k_2) < g(k_1)$. By virtue of the properties introduced in the remark 4.1, we know that there exists a finite time instant $k_{reach} \geq \bar{k}$ such that $g_r(k_{reach}) \equiv \bar{r}$, and $g(k_{reach}) \equiv \bar{r}$. According to the optimal solutions (4.18)-(4.19) at time k_1 is, by construction, a feasible solution for any $k \geq k_1$. Therefore, if $g_r(k_1) < g_r(k_2)$, and $g(k_1) < g(k_2)$ (strictly) we have an absurd saying that the solution at time k_1 is not feasible at k_2, i.e. $g_r(k_1) \notin \mathcal{V}(x(k_2))$, and $g(k_1) \notin \mathcal{V}(x(k_2))$. As a consequence $g_r(k)$, and $g(k)$ must have a MN-D behavior. The case $g_r(\bar{k}) > \bar{r}$, and $g(\bar{k}) > \bar{r}$, can be addressed with the same arguments and, in this case, the non-increasing behavior of both CGs can be proved. □

Such a setpoint attack detection rule is here extended to take advantage of Assumption 4.2. In particular, if reference signal $r(k)$ is piecewise constant, then in the absence of

attack, the signal $g'_r(k)$ must also have a MN-D or MN-I behavior in each constant interval $[i\gamma, (i+1)\gamma)$ (see Remark 4.1). Therefore, the resulting setpoint attack anomaly detection rule is

$$\textit{Rule-2}(k,i)=\begin{cases} \text{anomaly}, & \textit{if} \begin{cases} g(k) \neq g'_r(k) \\ \quad \text{OR} \\ g'_r \text{ is neither MN-D nor MN-I} \\ \quad \text{for } k \in [i\gamma, (i+1)\gamma) \end{cases} \\ \text{normal}, & \textit{otherwise} \end{cases} \tag{4.20}$$

Remark 4.4. *The the detection rule (4.20) is robust against the disturbances (4.2) and, as a consequence, false positive occurrences are avoided.* □

By collecting the above two rules, the robust detector module in Figure. 4.2 exploits the following logic rules:

$$\textit{Detector}(k,i)=\begin{cases} \text{anomaly}, & \textit{if} \begin{cases} \textit{Rule-1}(k) = \text{anomaly} \\ \quad \text{OR} \\ \textit{Rule-2}(k,i) = \text{anomaly} \end{cases} \\ \text{normal}, & \textit{otherwise} \end{cases} \tag{4.21}$$

4.5.2 Emergency Controller

The emergency controller must be designed to replace the networked controller when a cyber-attack makes the communication channels unreliable. In particular, the proposed emergency controller is a state-feedback controller

$$u_\varepsilon(k) = \varepsilon(x_p(k)), \quad \varepsilon : \mathcal{X}_\varepsilon \subset \mathbb{R}^{n_p} \to \mathcal{U}_\varepsilon \subset \mathbb{R}^m \tag{4.22}$$

which satisfies the following requirements:

- *(R_1)* The controller's domain covers the networked controller's DoA and fulfills the constraints (4.3)-(4.4)

$$\mathcal{X}_c \subseteq \mathcal{X}_\varepsilon \subseteq \mathcal{X}_p, \quad \mathcal{U}_\varepsilon \subseteq \mathcal{U} \tag{4.23}$$

- *(R_2)* The controller is capable of confining, in a finite number of steps, the plant's state trajectory into a RCI terminal emergency region, namely $\mathcal{X}_e$, such that

$$\mathcal{X}_e \subset \mathcal{X}_c \tag{4.24}$$

Remark 4.5. *It is important to remark that the reference signal $r(k)$ is not available on the plant side. As a consequence, the emergency controller cannot solve the reference tracking problem as done by the networked controller. The only objective of the emergency controller is to preserve the plant's safety until an attack-free scenario is recovered. Moreover, to ensure that the emergency controller can be activated from any admissible state $x_p(k) \in \mathcal{X}_c$, its domain $\mathcal{X}_\epsilon$ must contain $\mathcal{X}_c$. Finally, the condition $\mathcal{X}_e \subset \mathcal{X}_c$ is instrumental in ensuring that when the cyber-attack is ended, the networked controller can be reactivated. Please refer to Figure. 4.3 for an illustration of the state containment conditions prescribed by (4.23)-(4.24).*

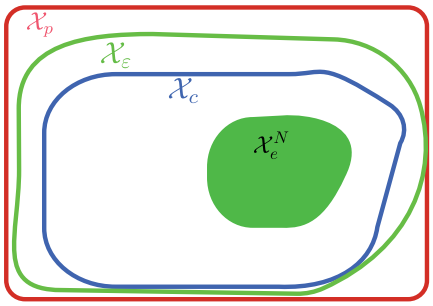


Figure 4.3: Safety state constraint ($\mathcal{X}_p$), networked controller's DoA ($\mathcal{X}_c$), emergency controller's DoA ($\mathcal{X}_\varepsilon$), and emergency controller RCI terminal region ($\mathcal{X}_e$).

The next proposition defines the set of states $N-$step safe regardless of any admissible disturbance realization (4.2) and FDI attack (4.14). Such a condition will be leveraged to build the emergency controller.

Definition 4.2. *A state $x_p(k)$ is said $N-$step attack safe if, in the worst-case scenario, the state trajectory starting from $x_p(k)$ is confined within $\mathcal{X}_c$, i.e.,*

$$x_p(k+i) \in \mathcal{X}_c, \forall\, i = 0, \ldots, N-1 \tag{4.25}$$

□

Proposition 4.2. *Consider the plant evolution (4.15), the state and input constraints (4.3)-(4.4), the networked controller's domain of attraction $\mathcal{X}_c$. Then the plant's state trajectory evolution is guaranteed to be $N-$step attack-safe, regardless of any FDI attack (4.14) and disturbance occurrences (4.2) if the current state vector $x_p(k)$ belongs to a set $\mathcal{X}_e^N$ satisfying the following condition*

$$\begin{cases} A^{N-1}\mathcal{X}_e^N \subseteq \tilde{\mathcal{X}}_c \\ A\mathcal{X}_e^N \subseteq \tilde{\mathcal{X}}_c \end{cases} \tag{4.26}$$

where

$$\tilde{\mathcal{X}}_c = \mathcal{X}_c \ominus \left(\sum_{j=0}^{N-1} A^i (B\mathcal{U} \oplus B_d\mathcal{D}) \right)$$

Proof: In an attack-free scenario ($u(k) \equiv u'(k)$, $x(k) \equiv x'(k)$, and $g_r'(k) = g_r(k)$), despite any admissible disturbance realization (4.2), the networked controller ensures, by construction, that all the constraints are satisfied (4.3)-(4.4) and that $x_p(k) \in \mathcal{X}_c$, $\forall\, k$.

The presence of FDI attacks (4.14) can directly (attacking the control input) and/or indirectly (attacking the state measurement) change the control inputs received by the plant, i.e., $u'(k) \neq u(k)$. Let $x_p(k)$ be the current plant state. Then, the $i-$step ahead plant evolution (4.15) is described as follows

$$\hat{x}_p(k+i) = A^i x_p(k) + \sum_{j=0}^{i-1} A^{i-1-j}(B\hat{u}'(k+j) + B_d\hat{d}(k+j)) \tag{4.27}$$

where $\hat{x}_p(\cdot)$, $\hat{u}'(\cdot)$ and $\hat{d}(\cdot)$ are the predicted states, command inputs imposed by the

attacker and disturbances, respectively. In principle, the attacker can arbitrarily decide $u'(k)$, however, due to physical actuation constraints (4.3), the control input applied to the plant is saturated, i.e., $u'(k) \in \mathcal{U}$. Therefore, the state $x_p(k)$ is safe for $N-$steps, regardless of any admissible disturbance realization (4.2) and FDI attack (4.14), if the predicted system evolution for $N-$steps remains confined within the networked controller's DoA (4.12) for any admissible input imposed by the attacker, i.e.,

$$\begin{gathered} \forall i \in \{0, \ldots, N-1\}, \forall \hat{d}(k+i) \in \mathcal{D}, \forall \hat{u}'(k+i) \in \mathcal{U} \rightarrow \\ \hat{x}_p(k+i) \in \mathcal{X}_c \end{gathered} \tag{4.28}$$

Therefore, by considering the worst-case scenario and by resorting to the Minkowsky/Pontryagin set-sum, we can rewrite the condition (4.28) as follows

$$\forall\, i \in \{0, \ldots, N-1\} \rightarrow A^i x_p(k) + \sum_{j=0}^{i-1} A^i (B\mathcal{U} \oplus B_d \mathcal{D}) \subseteq \mathcal{X}_c \tag{4.29}$$

By resorting to the Minkowsky set-difference operator, we obtain

$$\forall\, i \in \{0, \ldots, N-1\} \;\; A^i x_p(k) \subseteq \tilde{\mathcal{X}}_c \tag{4.30}$$

where

$$\tilde{\mathcal{X}}_c = \mathcal{X}_c \ominus \left(\sum_{j=0}^{N-1} A^j (B\mathcal{U} \oplus B_d \mathcal{D}) \right).$$

Finally, by exploiting the set inclusion property of the sets $A^i \mathcal{X}_e^N$, $\forall\, i$ [100, Remark 4.1], the condition (4.30) can be further simplified as follows:

$$\begin{cases} A^{N-1} \mathcal{X}_e^N \subseteq \tilde{\mathcal{X}}_c \\ A \mathcal{X}_e^N \subseteq \tilde{\mathcal{X}}_c \end{cases} \tag{4.31}$$

concluding the proof. □

Remark 4.6. *As long as the computation of $\mathcal{X}_e^N$ is concerned, please note that we can*

write $\mathcal{X}_e^N$ *as the following set*

$$\mathcal{X}_e^N = \left\{ x_p \in \mathbb{R}^{n_p} : \begin{array}{c} A^{N-1}x_p \subseteq \tilde{\mathcal{X}}_c \\ Ax_p \subseteq \tilde{\mathcal{X}}_c \end{array} \right\} \tag{4.32}$$

If the face representation of the polyhedron $\tilde{\mathcal{X}}_c$ *is considered*

$$\tilde{\mathcal{X}}_c : \tilde{T}_c x_p \leq \tilde{b}_c$$

with $\tilde{T}_c$*, and* $\tilde{b}_c$ *a matrix and a vector of compatible size, we can re-write (4.32) as*

$$\mathcal{X}_e^N = \left\{ x_p \in \mathbb{R}^{n_p} : \begin{array}{c} \tilde{T}_c A^{N-1}x_p \leq \tilde{b}_c \\ \tilde{T}_c Ax_p \leq \tilde{b}_c \end{array} \right\} \tag{4.33}$$

Then, $\mathcal{X}_e^N$ *is given by the following polyhedral set*

$$\mathcal{X}_e^N : \ T_e^N x_p \leq b_e^N \tag{4.34}$$

where

$$T_e^N = \begin{bmatrix} \tilde{T}_c A^{N-1} \\ \tilde{T}_c A \end{bmatrix}, \quad b_e^N = \begin{bmatrix} \tilde{b}_c \\ \tilde{b}_c \end{bmatrix}$$

□

Given the $N-$step attack safe region $\mathcal{X}_e^N$, satisfying (4.26), the emergency controller's is offline built computing a family or robust of robust one-step controllable set $\{\mathcal{T}_l\}_{l=0}^{L}$ [75] according to the following definition, which is recursively applied until the condition R_1, i.e., $\mathcal{X}_\varepsilon \supseteq \mathcal{X}_c$, is fulfilled:

$$\begin{aligned} \mathcal{T}_0 &:= \mathcal{X}_e^N \\ \mathcal{T}_l &= \{x_p \in \mathbb{R}^{n_p} : \exists u \in \mathcal{U} \text{ s.t. } Ax_p + Bu \in \tilde{\mathcal{T}}_{l-1}\},\ l \geq 1 \end{aligned} \tag{4.35}$$

where $\tilde{\mathcal{T}}_l := \mathcal{T}_l \ominus B_d\mathcal{D}$.

Given $\mathcal{X}_\varepsilon := \bigcup_{l=0}^{L} \mathcal{T}_l \supseteq \mathcal{X}_c$, the emergency controller (4.22) actions are online computed by means of the following receding horizon algorithm [79]:

Emergency Controller (EC) Algorithm - $\forall k$

Initialization: $\{\mathcal{T}\}_{l=0}^{L}$

Input: $x_p(k)$

Output: $u_\varepsilon(k)$

1: Find the smallest set index $l(k)$ containing $x_p(k)$, i.e.,

$$l(k) := \min_{1\leq l\leq L} \{l : \; x_p(k) \in \mathcal{T}_l\} \tag{4.36}$$

2: Solve the following convex optimization problem:

$$\begin{aligned} u_\varepsilon(k) = \arg\min_u J(x_p(k), u) \quad s.t. \\ Ax_p(k) + Bu \in \tilde{\mathcal{T}}_{l(k)-1}, \; u \in \mathcal{U} \end{aligned} \tag{4.37}$$

3: $k \leftarrow k+1$ goto Step 1

---

where $J(x_p(k), u)$ is any convex cost function of interest.

Remark 4.7. *Please note that the recursion (4.35) defines a family of $L \geq 1$ robust one-step controllable sets $\{\mathcal{T}_l\}_{l=0}^{L}$ that guarantees the existence of an emergency controller whose DoA is [75]*

$$\mathcal{X}_\varepsilon := \bigcup_{l=0}^{L} \mathcal{T}_l$$

Indeed, for any state $x_p(k) \in \mathcal{T}_l$, $1 \leq l \leq L$, there exists (by construction) an admissible control input capable of robustly steering the one-step evolution within the predecessor of the current set, i.e., $x_p(k+1) \in \mathcal{T}_{l-1}$. Therefore, in at most L steps, any state $x_p(k) \in \bigcup_{l=0}^{L}\{\mathcal{T}_l\}$, can be steered in the emergency region $\mathcal{X}_e^N$ (satisfying the emergency controller requirement R_2). The latter ensures that the convex optimization problem (4.37) always admits a solution in polynomial time and that emergency controller algorithm enjoys recursive feasibility. Finally, the cost function $J(x_p(k), u)$ can be arbitrarily chosen and it can be used to penalize any desired convex combination of speed convergence and control effort. □

4.5.3 Smart Actuator

The smart actuator must ensure that regardless of the attacker actions (4.14) and detector (4.21) performance, the *safety* of the system (see *Definition* 4.1) is guaranteed. Therefore, it is responsible of detecting potential plant's safety violation occurrences and decide to apply either the command received through the networked ($u'(k)$) or the one computed by the emergency controller ($u_\varepsilon(k)$).

To this end, the actuator exploits the robust one-step reachable set, namely $\mathcal{X}^+(x_p, u')$, to detect any possible safety constraint violation at least one step before their occurrences. In particular, given the current plant state $x_p(k)$ and the received control input $u'(k)$, the actuator decides the control signal to apply as follows

$$u_{SA}(k) = \begin{cases} u'(k) & \text{if } \mathcal{X}^+(x_p(k), u'(k)) \subseteq \mathcal{X}_c \\ u_\varepsilon(k) & \text{otherwise} \end{cases} \tag{4.38}$$

Remark 4.8. *The concept of $N-$step attack safe region (Definition 4.2), with $N = 1$, i.e.,*

$$\mathcal{X}_e^1 = \left\{ x_p \in \mathbb{R}^{n_p} : \ \tilde{T}_c A x_p \leq \tilde{\mathcal{X}}_c \right\} \tag{4.39}$$

can be used instead of $\mathcal{X}^+(x_p(k), u'(k))$ in (4.38) to detect safety constraint violation [68].

However, since

$$\mathcal{X}^{+}(x_p(k), u'(k)) \subseteq \mathcal{X}_c \Rightarrow x_p(k) \in \mathcal{X}_e^1$$

the condition used in (4.38) is less conservative.

Moreover, the smart actuator takes care of possible intermittent FDI attacks aiming to produce instability by triggering a frequent switch among the two above control laws [101]. To avoid such a possibility, if the emergency controller is used at the time k, it will be applied for the next successive $L-$steps to ensure that the $N-$step safe region $\mathcal{X}_e^N$ is reached. Finally, if an attack is detected by the detector (4.21), then the consequent network disconnection (Assumption 4.3) will trigger the actuator ($u'(k) = \emptyset$) to directly activate the emergency controller.

Remark 4.9. *Please note that the used attack detector (4.21) is robust against false positive (see Remarks 4.3-4.4), Therefore, in the absence of attacks, the emergency controller will never be triggered.*

The following pseudo algorithm summarizes the operations performed by the smart actuator:

Smart Actuator (SA) Algorithm - $\forall k$

Initialization: $anomaly = 0$, $counter = 0$.

Input: $u'(k)$, $u_\varepsilon(k)$, L

Output: $u_{SA}(k)$

1: Set $check_1 = true$ if $(u'(k) \neq \emptyset \ \ \& \ \ u'(k) \in \mathcal{U})$

2: Set $check_2 = true$ if $\mathcal{X}^{+}(x_p(k), u'(k)) \subseteq \mathcal{X}_c$

3: Set $check_3 = true$ if $anomaly == 0$

4: **if** $check_1 == true \& check_2 == true \& check_3 == true$ **then**

5: $anomaly = 0$

6: $u_{SA}(k) = u'(k)$ ▷ (Networked Controller)

7: **else**

8: $u_{SA}(k) = u_{\varepsilon}(k)$ ▷ (Emergency Controller)

9: $counter = counter + 1$

10: **if** $anomaly = 0$ **then** $anomaly = 1$

11: **else** ▷ (Hold $u_{\varepsilon}(k)$ for L steps)

12: **if** $counter == L$ **then**

13: $counter == 0$, $anomaly = 0$;

14: **end if**

15: **end if**

16: **end if**

Proposition 4.3. *Given the constrained plant model (4.1)-(4.4) and the networked controller domain of attraction $\mathcal{X}_c$, the proposed control architecture (Figure. 4.2) fulfills the objectives* ***(CH4-O1)-(CH4-O2)***

Proof: The proposition can be proved by collecting all the above developments: detector (4.21), emergency controller algorithm (EC), and the smart actuator algorithm (SA).

(CH4-O1) - The proposed detector (4.21) is able to detect anomalies caused by FDI attacks in the C-P, P-C and C-C channels. Moreover, since once an attack is detected then the communication channels between the plant and the controller are interrupted (*Assumption* 4.3) then the resulting $u'(k) = \emptyset$ received by the smart actuator will trigger the emergency controller (see Steps 1 and 4 of the SA algorithm) stopping the attacker actions. Moreover, the actuator checks in steps 1-2 of the SA algorithm ensure that, regardless of the detector performance, any attack is discovered and the emergency controller

activated at least one-step before the plant could reach the unsafe condition $x_p(k) \notin \mathcal{X}_c$ (see *Remark* 4.8).

(CH4-O2) - In the absence of cyber-attacks, the networked controller ensures that plant state trajectory is, by design, confined in $\mathcal{X}_c$ and that the constraints (4.3)-(4.4) are fulfilled. In the presence of attacks, the emergency satisfies, by design, the requirements R_1 in (4.23) and R_2 in (4.24) ensuring constraint satisfaction and state trajectory Uniform ultimate boundedness in the emergency $N-$step attack safe region $\mathcal{X}_e^N \subset \mathcal{X}_c$. Therefore, if an attack-free scenario is recovered, and L steps are elapsed from the activation of the emergency controller (step 13 of the SA algorithm), then the networked tracking controller can be re-activated. □

4.6 Simulation Example

In this section, the industrial Continuous-Stirred Tank Reactor (CSTR) system, described in section 3.3.9, is considered, see Figure. 3.12.

The CSTR linearized dynamics are described by (3.46). Moreover, the plant bounded disturbance set is $\mathcal{D} = \{d : [-0.01, -0.8]^T \leq d \leq [0.01, 0.8]^T\}$, and the following state and input constraints are prescribed:

$$\begin{aligned} \mathcal{U} &= \{[T_C,\, C_{Ai}]^T \in \mathbb{R}^2 :\ -2 \leq T_C \leq 2,\ \ -10 \leq C_{Ai} \leq 10\} \\ \mathcal{X}_p &= \{[C_A,\, T]^T \in \mathbb{R}^2 :\ -10 \leq C_A \leq 10,\ \ \ -30 \leq T \leq 30\} \end{aligned} \tag{4.40}$$

The networked CG tracking controller has been implemented using, as primal controller, an LQI controller, and with the CG configuration $k_0 = 280$ and $\delta = 10^{-5}$. The resulting networked controller has a domain of attraction $\mathcal{X}_c \subset \mathcal{X}_p$ shown in Figure. 4.9 (dashed-line polyhedron).

The emergency controller has been designed as follows. First, the state space region $N = 9-$step safe, i.e., $\mathcal{X}_e^9$, has been computed as in (4.33) (green region in Figure. 4.9).

Then, the emergency controller is offline built using the recursion (4.35) under the conditions (4.23)-(4.24). Specifically, $L = 17$ robust one-step controllable sets, $\{\mathcal{T}_l\}_{l=1}^{17}$ have been computed (black solid-line polyhedral regions in Figure. 4.9).

In what follows two attack scenarios are simulated to better explain the operations of the proposed control strategy:

- A FDI attack on the setpoint signal $g_r(k)$ transmitted by the command center (attack on the C-C channel)
- An undetectable covert attack [26] on the control signal $u(k)$ and plant measurements $x_p(k)$ (attack on the C-P and P-C channels)

FDI attack on the Setpoint Signal: In this experiment, the reference signal $r(k)$ is

$$r(k) = \begin{cases} [0, 0]^T & 0 \leq k \leq 69 \\ [-0.4, 10]^T & 70 \leq k \leq 209 \\ [1, 17]^T & k \geq 210 \end{cases} \tag{4.41}$$

Moreover, the attacker for $150 \leq k \leq 240$ replaces the legitimate reference signal $g_r(k)$ with a desired constant vector, i.e.,

$$g_r'(k) = \begin{cases} [-1, -17]^T & 150 \leq k \leq 240 \\ g_r(k) & otherwise \end{cases} \tag{4.42}$$

The simulation results pertaining to this case are shown in Figures. 4.4-4.5. First in Figure. 4.4, it is possible to appreciate that such attack is instantaneously detected by the detector (4.21). Indeed, at $k = 150$ the reference signal received by the CG is not an admissible, i.e., $[g_r'^T(150), x^T(150)]^T \notin \mathcal{Z}$. In particular, the output of the CG, $g(150) = [-0.49; -2.73]^T$, is different from $g_r'(150) = [-1, -17]^T$ triggering the detection rule $Rule-2(k, i)$ (see Figure. 4.5). Moreover, the networked controller disconnects the C-P channel

(assumption 4.3), i.e., $u'(k) = \emptyset$, triggering the Smart Actuator to activate the Emergency Controller (see Steps 1 and 7 of the SA-algorithm). As a consequence, the state trajectory is steered towards the 9−step attack safe region $\mathcal{X}_e^9$ without any constraint violation (Figures. 4.6 and 4.5) and confined in it until the attack is terminated. At $k = 241$ the C-P channel is re-established, the Smart Actuator re-activates the networked controller, and the plant start tracking-back the current reference signal (Figure. 4.5).

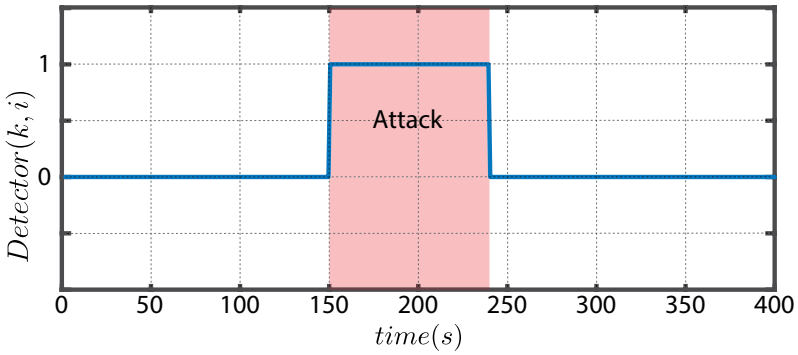


Figure 4.4: Detector Output in the presence of the FDI attack on the setpoint signal.

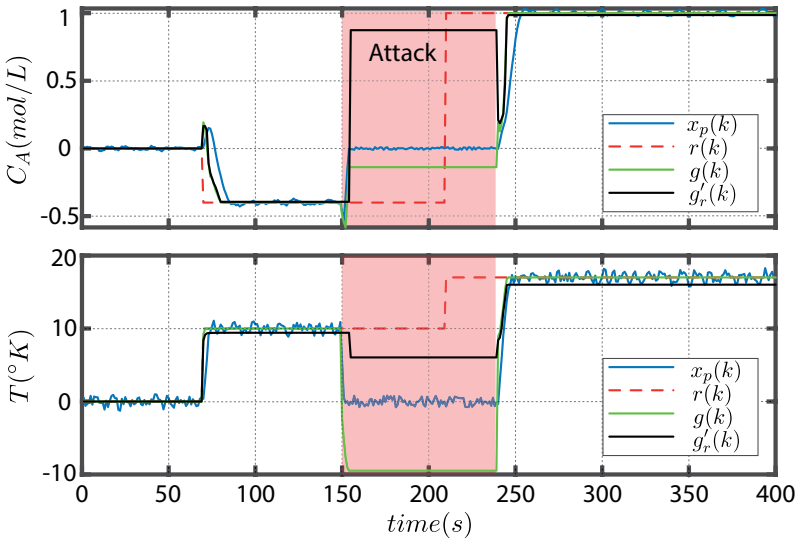


Figure 4.5: Plant's states evolution in the presence of the FDI attack on the setpoint signal.

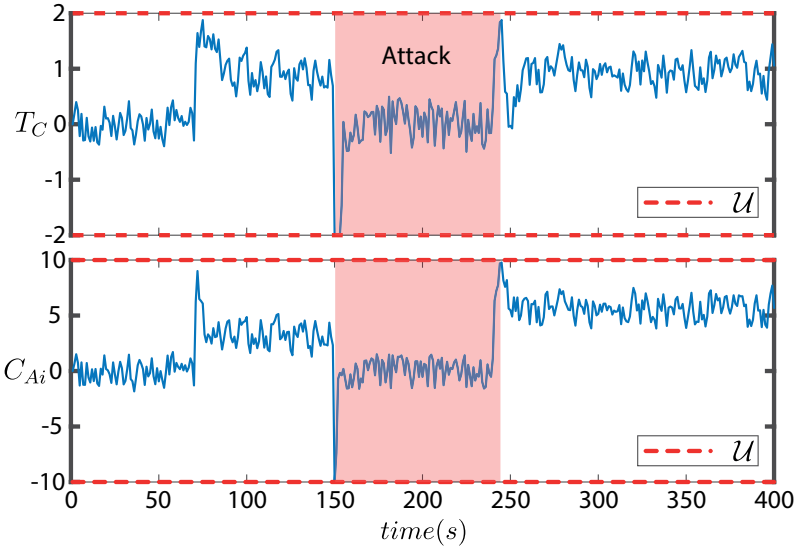


Figure 4.6: Control signals in the presence of the FDI attack on the setpoint signal.

Undetectable covert attack: In this experiment, the reference signal $r(k)$ is

$$r(k) = \begin{cases} [0,0]^T & 0 \leq k \leq 69 \\ [-3,5]^T & 70 \leq k \leq 209 \\ [1,17]^T & k \geq 210 \end{cases} \tag{4.43}$$

while the attacker performs an undetectable covert attack on the C-P and P-C channels. This simulation aims to show that even in the presence of stealthy attacks, the proposed control architecture still ensures the plant's safety. The covert attack aims to bring the state trajectory in the unsafe configuration $x_p = [11, 25]^T \notin \mathcal{X}_p$ while avoiding detection. To this end, the following FDIs on the actuation and measurement channels is performed

$$u'(k) = \begin{cases} u(k) + u^a(k) & 150 \leq k \leq 240 \\ u(k) & otherwise \end{cases}. \tag{4.44}$$

$$x'_p(k) = \begin{cases} x_p^a(k) & 150 \leq k \leq 240 \\ x_p(k) & otherwise \end{cases}. \tag{4.45}$$

with

$$u^a(k) = sat\left(-K\left(x(k) - \begin{bmatrix}11\\25\end{bmatrix}\right) + \begin{bmatrix}-1.4591\\9.4475\end{bmatrix}\right) - u(k) \tag{4.46}$$

and where $sat(\cdot)$ is a function that saturates the control input vector inside the constraints $\mathcal{U}$,

$$K = \begin{bmatrix}-4.89 & 0.45\\0.86 & 1.96\end{bmatrix} \tag{4.47}$$

the stabilizing control gain used by the attacker, and

$$x_p^a(k) = A^{k-1}x_p(150) + \sum_{j=0}^{k-1}(A^j B u(k-1-j)). \tag{4.48}$$

Remark 4.10. *Please note the attack on the measurement channel replaces the actual output to hide the attack on the actuation channel, see [9] for further details about the considered covert attack.*

The results pertaining this simulation are collected in Figures 4.7-4.9. First, it is important to remark that the designed covert attack cannot be detected by the proposed detector (4.21). Therefore, for $150 \leq k < 210\,\text{sec}$ the attacker is able to steer the state trajectory away from the desired reference signal (see Figures. 4.7 and 4.9 (blue solid-line)). Nevertheless, at $k = 210$, on the plant side, the Smart Actuator detects a potential safety risk for the plant, i.e., the one-step evolution of the system can exit the controller domain $\mathcal{X}_c$, i.e., $\mathcal{X}^+(x_p(210), u'(210)) \notin \mathcal{X}_c$ and it activates the emergency controller (see Step 2 and 7 of the SA-algorithm, and Figure. 4.9 (red solid-line)) avoiding any constraint violation (Figures 4.7-4.8). At $k = 210$, $x_p(210) \in \mathcal{T}_{16}$, the EC algorithm is started and the state trajectory reaches $\mathcal{X}_e^9$ at $k = 226$. Moreover, according to the SA-algorithm, at $k = 227$ (after that $L = 17$ steps from the activation of the emergency controller are elapsed), the SA re-activate the networked controller. However, since the attack is still ongoing, the state trajectory starts moving away from the terminal region until at $k = 240$

when the attack ends. Finally, for $k > 240$ the networked controller is capable of recovering the normal tracking operations.

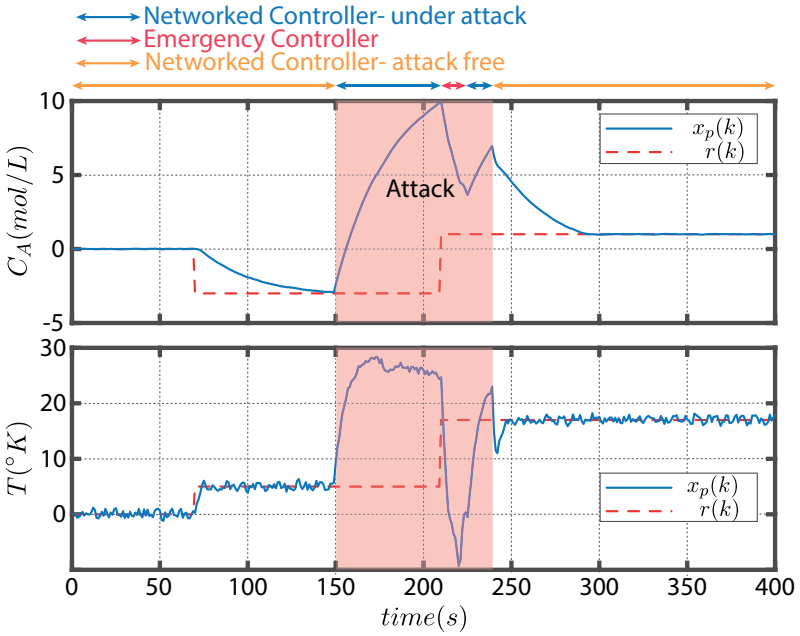


Figure 4.7: Plant's states evolution under the covert attack.

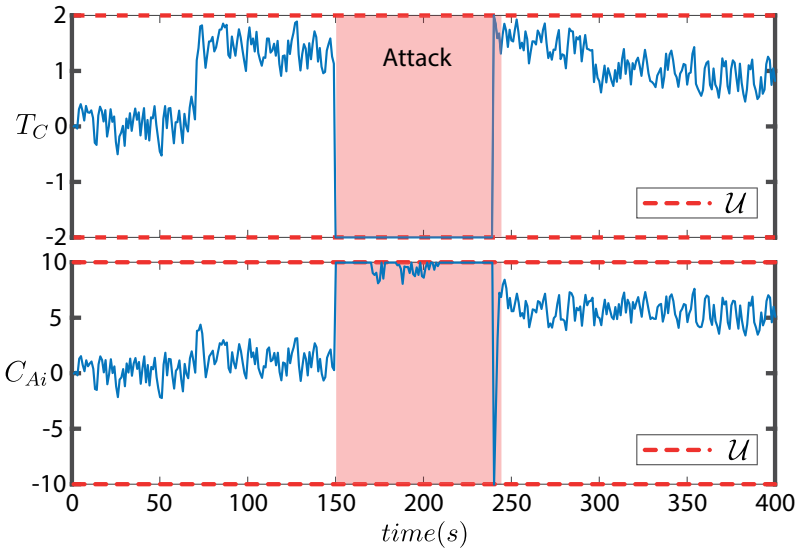


Figure 4.8: Control signals in the presence of the covert attack.

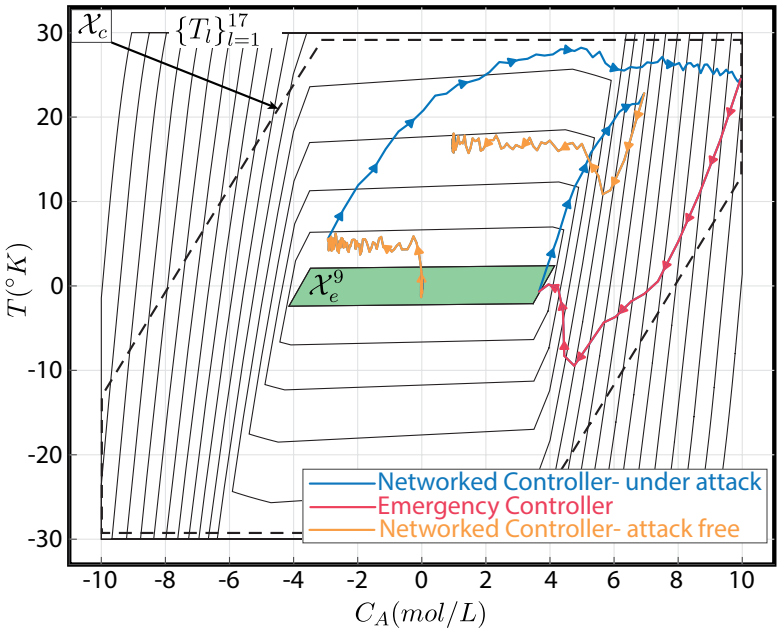


Figure 4.9: Networked controller domain $\mathcal{X}_c$, attack safe region $\mathcal{X}_e^9$, and state trajectory under the covert attack.

4.7 Conclusion

In this chapter, a novel networked control architecture has been proposed to ensure plant safety against a variety of cyber-attacks that can affect the communication channels in cyber-physical systems. This has been achieved by properly combining two main ingredients: (i) a detector module, local to the networked controller, capable of discovering the presence of a variety of FDI attacks on the control input, sensor measurement, and setpoint signals, (ii) a smart actuator module, local to the plant, capable of activating an ad-hoc designed emergency controller at least one-step before any plant safety constraint could be violated. By resorting to set-theoretic arguments, it is formally proved that plant safety is ensured regardless of the attacker's actions and detector performance. Simulation results obtained considering an industrial Continuous-Stirred Tank Reactor (CSTR) have been shown to testify the proposed approach effectiveness and validate the theoretical claims of the chapter.

Chapter 5

Reference tracking for cyber-physical systems under network attacks

5.1 Introduction

In this chapter, the safety and reference tracking control problems for Cyber-Physical Systems (CPSs) equipped with authenticated communication channels are addressed. In this class of CPSs, network attacks can break the feedback loop at two different points for an arbitrarily long period. In this scenario, we design a novel control architecture that, by taking a worst-case approach, aims to preserve the safety of the systems while minimizing, whenever possible, the tracking performance degradation. On the plant side, a local safety controller is designed to take care of attacks on the actuation channel. In particular, given a finite number of pre-determined admissible safe equilibrium points, this unit exploits a Voronoi partition of the state space and a family of dual-model set-theoretic model predictive controllers to safely confine, in a finite number of steps, the system into the closest robust control invariant region. On the other hand, on the controller side, the reference tracking controller operations are enhanced with an add-on module in charge of dealing with attack occurrences on the measurement channel. Specifically, by leveraging the Voronoi partition used on the plant's side and reachability arguments, the objective

of this unit is to reduce the performance loss by allowing a supervised system evolution until the best outcome in terms of tracking is achieved. The obtained theoretical results are proved and the solution's effectiveness is shown through a simulation example.

5.1.1 Contribution of the work

To the best of the our knowledge, no existing approaches look at the reference tracking problem for constrained systems under arbitrary attacks on both the actuation and measurement channels. Therefore, this chapter represents a first step towards addressing such an issue, and the peculiar capability of the proposed control framework can be summarized as follows: (i) the safety of the plant and post-attack recovery are guaranteed regardless of the attack actions and duration; (ii) the proposed solution is robust against the unknown but bounded process and measurement disturbances; (iii) under attacks, the proposed solution minimizes, whenever possible (according to a worst-case reachability analysis), the tracking performance degradation; (iv) the proposed solution consists of two add-on modules that can be added to existing networked infrastructure (i.e., there is no need to re-design the tracking controller to deal with the presence of attacks).

5.2 Networked Control System Setup

In this section, first, some preliminary definitions are provided, then, the considered networked control system setup is described.

Definition 5.1. *Given a point $p \in \mathbb{R}^s$ and a set $\mathcal{S} \subset \mathbb{R}^s$, the maximum distance between p and $\mathcal{S}$ is defined as*

$$d^{sup}(\mathcal{S}, p) \triangleq \sup\{\|p - s\|_2 : s \in \mathcal{S}\}.$$

5.2.1 Constrained Plant Model

We consider the class of discrete-time Linear Time-Invariant (LTI) systems subject to unknown but bounded process disturbances

$$x(k+1) = Ax(k) + Bu(k) + B_d d(k) \tag{5.1}$$

where $k \in \mathbb{Z}_+ = \{0, 1, \ldots\}$ is the discrete-time instant, $x(k) \in \mathbb{R}^n$ the state vector, $u(k) \in \mathbb{R}^m$ the control input vector, $d(k) \in \mathbb{R}^{n_d}$ a uniformly distributed process disturbance belonging to the compact set $\mathcal{D} \subset \mathbb{R}^{n_d}$ containing the origin

$$d(k) \in \mathcal{D},\ 0_{n_d} \in \mathcal{D}, \tag{5.2}$$

and A, B, B_d are the system matrices of appropriate dimensions. Moreover, the states and control inputs are constrained in compact subsets $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{U} \subset \mathbb{R}^m$, respectively, where

$$x(k) \in \mathcal{X},\ 0_n \in \mathcal{X}, \qquad u(k) \in \mathcal{U},\ 0_m \in \mathcal{U} \tag{5.3}$$

Definition 5.2. *Consider the plant model (5.1) under (5.2)-(5.3), and a set $\mathcal{T}_i \subset \mathcal{X}$. The set of states $\mathcal{T}_{i+1} \subset \mathcal{X}$ Robust One-Step Controllable (ROSC) to $\mathcal{T}_i$ is defined as [75]*

$$\mathcal{T}_{i+1} = \{x \in \mathcal{X},\ \exists u \in \mathcal{U} : Ax + Bu + B_d d \in \mathcal{T}_i,\ \forall d \in \mathcal{D}\} \tag{5.4}$$

We assume that the state vector can be either measured or estimated with a uniformly distributed and bounded measurement error $m(k) \in \mathcal{M} \subset \mathbb{R}^n$, where $\mathcal{M}$ is a compact set containing the origin, i.e.,

$$y(k) = x(k) + m(k), \quad m(k) \in \mathcal{M},\ 0_n \in \mathcal{M} \tag{5.5}$$

with $y(k) \in \mathbb{R}^n$ the measured vector. Given (5.1) and (5.5), it is possible to express the

uncertain one-step output evolution of the system as

$$y(k+1) = Ay(k) + Bu(k) + B_d d(k) - Am(k) + m(k+1) \tag{5.6}$$

and define the set of safe outputs, namely $\mathcal{Y} \subset \mathbb{R}^n$, as $\mathcal{Y} = \mathcal{X} \ominus (-\mathcal{M})$. Please note that if $y(k) \in \mathcal{Y} \rightarrow x(k) = y(k) - m(k) \in (\mathcal{X} \ominus (-\mathcal{M})) \oplus (-\mathcal{M}) \subseteq \mathcal{X}, \forall\, m(k) \in \mathcal{M}$.

5.2.2 Networked Tracking Controller

The plant is required to track a reference signal $r(k) \in \mathbb{R}^{n_r}$, $1 \leq n_r \leq n$. In particular, by defining $y^r(k) = Cx(k),\ y^r(k) \in \mathbb{R}^{n_r},\ C \in \mathbb{R}^{n_r \times n}$ as the vector required to track $r(k)$, we assume that a networked stabilizing tracking controller is available and its actions are generically modeled as

$$u(k) := \Phi(z_c, y^r(k), r(k)) \tag{5.7}$$

with $z_c \in \mathbb{R}^{n_c}$ the state vector of the controller, and $\Phi(\cdot,\cdot,\cdot)$ the control logic. The output subspace

$$\mathcal{Y}_c \subseteq \mathcal{Y} \tag{5.8}$$

from which the controller solves the reference tracking problem while avoiding constraints (5.3) violation is hereafter denoted as the Tracking Domain of Attraction (T-DoA). In what follows, $\bar{y}_{r(k_1)} \in \mathbb{R}^n$ denotes the equilibrium output for the disturbance-free model of (5.6) for $r(k) = r(k_1)\ \forall\, k \geq k_1$.

5.2.3 Communication Channels, Cyber-Attacks, Safety

The communication channels between the plant and the controller are authenticated, i.e., a Message Authentication Code (MAC) [102] 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. Moreover, if an attack is present on a channel, then the received packets are considered unreliable and dropped.

We assume that cyber-attacks on the actuation and measurement channels can start at asynchronous times and have an arbitrarily long duration.

Remark 5.1. *The authors are aware that there is relevant literature that focuses on the attack detection problem using control-theoretical tools [10, 103]. Traditionally, such mechanisms are used when authenticated channels are not possible/supported or as a second security layer. Moreover, such solutions might not guarantee instantaneous attack detection [34]. Nevertheless, in the last years, in the cyber and network security communities, different lightweight encrypted and authenticated communications schemes have been designed to be computationally low-demanding, so being affordable in different CPS applications. To this end, particularly interesting is the set of benchmarks provided by the wolfSSL library*[1]*, showing how different encryption schemes (see e.g., AES Galois/Counter Mode (GCM) [104]) perform on different hardware; As an example, in [105, Table 1], it has been shown that for a typical quadruple-tank control system benchmark, the time required by the AES-GCM authenticated encryption scheme (when implemented on a standard Intel computed) is in the order of microseconds. Therefore, it is expected that encrypted communications and MAC to be widely available security features for the future generations of CPSs. The latter is also supported by the fact that encrypted control schemes are receiving increasing attention by the control community, see, e.g., the recent contributions [106–109] and references therein. As a consequence, the rest of the paper is developed by considering a setup where MAC is available. Nevertheless, for the sake of completeness, in the section 5.4.3, it is shown how the proposed solution can be adapted to a case where MAC is not available, and the attack detection mechanism introduces a detection delay.*

[1] https://www.wolfssl.com/docs/benchmarks/

5.3 Problem Formulation

In the considered networked architecture, we assume that only the networked controller is aware of the reference. Therefore, since $r(k)$ is unknown on the plant side, 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 cyber-attacks' impact on the system performance (i.e., minimizing the tracking performance degradation). More formally, the problem of interest can be stated as follows:

Given the plant model (5.1)-(5.5), the networked tracking controller (5.7) and its T-DoA (5.8), design a control architecture capable of (i) guaranteeing the plant's safety under cyber-attacks of arbitrary duration while (ii) reducing the tracking performance degradation and ensuring recovery.

5.4 Proposed Solution

To motivate the proposed solution, first the three admissible attack scenarios are analyzed:

- S_1: The cyber-attack affects only the actuation channel. Therefore $u(k)$ is invalid, while $y(k)$ is valid;
- S_2: The cyber-attack affects only the measurement channel. Therefore $y(k)$ is invalid, while $u(k)$ is valid;
- S_3: A cyber-attack affects both communication channels (synchronously or asynchronously). Therefore, both $y(k)$ and $u(k)$ are invalid starting from possible different time instants.

A possible safety preserving control solution for the above attack scenarios could be obtained by adopting the approach in [67], where a control architecture is designed considering the worst-case scenario, i.e., S_3. Therefore, as soon as an attack is detected (in either

channel), the communications are all interrupted, thus neglecting the consequences on the current reference tracking task. Such a solution, although effective to ensure the safety in $S_1 - S_3$, results to be over-conservative. In the proposed solution, we argue that lower tracking performance degradation can be obtained if *ad-hoc* actions are taken to explicitly deal with attacks on the actuation and measurement channels. To this end, we enhance the networked control scheme in Figure. 5.1 with two add-on modules:

- A *safety controller*, local to the plant, in charge of guaranteeing the plant safety when $u(k)$ is invalid.
- A *tracking supervisor*, local to the tracking controller, responsible for minimizing the tracking performance loss while ensuring safety when $y(k)$ is invalid.

In the proposed solution, we consider a setup where the existing tracking controller (5.7) cannot be re-designed and that a preview of the future reference signal is not available.

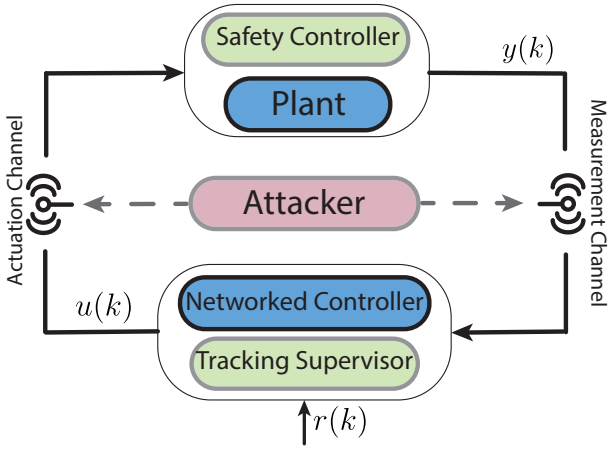


Figure 5.1: Proposed Control Architecture

5.4.1 Safety Controller (SC)

Since the Safety Controller (SC) does not have access to the reference signal $r(k)$, its objective is to preserve the plant safety (during the attack), and to guarantee performance recovery (when the attack is terminated). The latter translates into the following

requirements. Let

$$u^{sc}(k) = \eta_{sc}(y(k)), \quad \eta_{sc} : \mathcal{Y}_{sc} \subset \mathbb{R}^n \to \mathcal{U}_{sc} \subseteq \mathbb{R}^m \tag{5.9}$$

be the SC controller's logic with domain of attraction (DoA) $\mathcal{Y}_{sc}$. Then $\eta(\cdot, \cdot)$ must be designed such that:

$$\mathcal{Y}_c = \mathcal{Y}_{sc}, \quad \mathcal{Y}_{sc} \text{ is RCI}, \quad \mathcal{U}_{sc} \subseteq \mathcal{U} \tag{5.10}$$

where $\mathcal{Y}_c = \mathcal{Y}_{sc}$ (see Figure. 5.2) ensures that SC can replace the tracking controller from any admissible initial condition $y(k) \in \mathcal{Y}_{sc}$. Moreover, the conditions "$\mathcal{Y}_{sc}$ is RCI" and "$\mathcal{U}_{sc} \subseteq \mathcal{U}$" guarantee that the controller complies with the constraints (5.3) and that the performance recovery is ensured, i.e., the tracking controller can be enabled at any time without constraints violations.

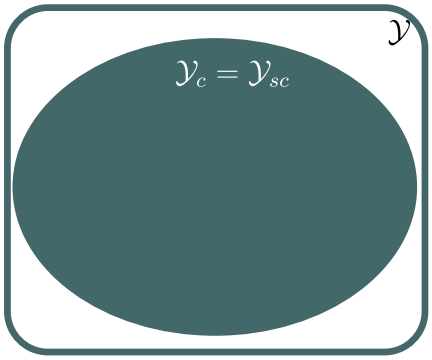


Figure 5.2: Safety controller's domain of attraction

To design such a controller we first consider a set of $L \geq 1$ predefined equilibrium pairs (x_e^l, u_e^l), $l \in \mathcal{L} := \{1, \ldots, L\}$ for the output disturbance-free model of (5.6). Moreover, a pair (y_e^l, u_e^l) is considered admissible/safe only if there exists a feedback controller with gain $K \in \mathbb{R}^{m \times n}$,

$$u^l(k) = K_l(y(k) - y_e^l) + u_e^l \tag{5.11}$$

such that the associated minimal RCI set [81] for (5.6), namely $\mathcal{T}_0^l \in \mathbb{R}^n$, is contained in the T-DoA, i.e., $\mathcal{T}_0^l \subseteq \mathcal{Y}_c$.

Note that the L controllers (5.11) and associated RCI sets $\mathcal{T}_0^l$, $l \in \mathcal{L}$ do not guarantee that the requirements (5.10) are fulfilled, i.e., it might exist $y \in \mathcal{Y}_c$ such that $y \notin \bigcup_{l=1}^{L} \mathcal{T}_0^l$. Therefore, to ensure compliance with (5.10), the following strategy is adopted.

A Voronoi partition of $\mathcal{Y}_c$ is created (see Figure. 5.3 as an example of partition) considering as generators (Voronoi seeds) the equilibrium points $\{y_e^l\}_{l=1}^{L}$. Therefore, a family of polyhedral regions $\{\mathcal{V}_l\}_{l=1}^{L}$ enjoying the following properties is obtained:

$$\mathcal{V}_l = \{y \in \mathcal{Y}_c : \|y - y_e^l\|_2 \leq \|y - y_e^j\|_2, \forall j \neq l, j \in \mathcal{L}\} \tag{5.12}$$

$$\bigcup_{l=1}^{L} \mathcal{V}_l = \mathcal{Y}_c. \tag{5.13}$$

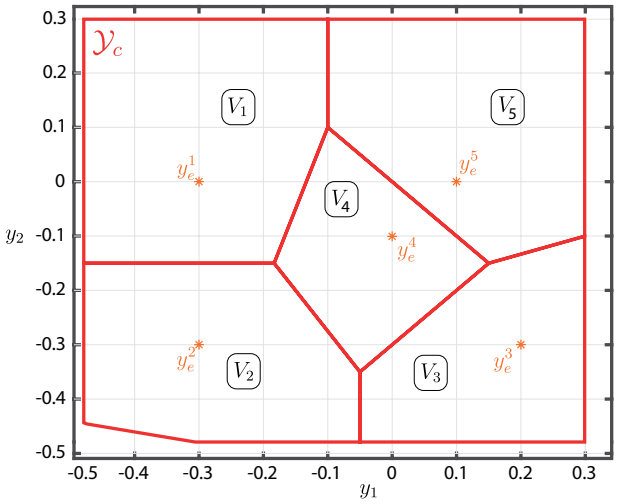


Figure 5.3: Voronoi partition for five equilibrium points

Then, by resorting to a dual-mode set-theoretic MPC paradigm [79], we enlarge the DoA of each $l-th$ controller (5.11) (i.e., $\mathcal{T}_0^l$) to cover the associated Voronoi partition $\mathcal{V}_l$ (see Figure. 5.4 as an example). To this end, a family of robustly controllable sets is recursively built by adapting the definition of ROSC sets (5.4) to the one-step output

evolution model (5.6) under (5.2)-(5.5):

$$\begin{aligned}\mathcal{T}_i^l = \; & \{y \in \mathcal{V}_l : \exists u \in \mathcal{U} : Ay + Bu + B_d d - Am_1 + m_2 \in \mathcal{T}_{i-1}^l, \\ & \forall d \in \mathcal{D},\, m_1, m_2 \in \mathcal{M}\} \\ = \; & \{y \in \mathcal{V}_l : \exists u \in \mathcal{U} : Ay + Bu \in \tilde{\mathcal{T}}_{i-1}^l\}\end{aligned} \tag{5.14}$$

with $\tilde{\mathcal{T}}_i^l = \mathcal{T}_i^l \ominus (B_d\mathcal{D} \oplus (-A\mathcal{M}) \oplus \mathcal{M})$. In particular, families $\{\mathcal{T}_i^l\}_{i=1}^{N_l}$, $l \in \mathcal{L}$ of $N_l \geq 0$ of ROSC sets are built, with N_l satisfying the termination condition

$$\bigcup_{i=1}^{N_l} \{\mathcal{T}_i^l\} = \mathcal{V}_l, \quad \forall\, l = 1, \ldots, L \tag{5.15}$$

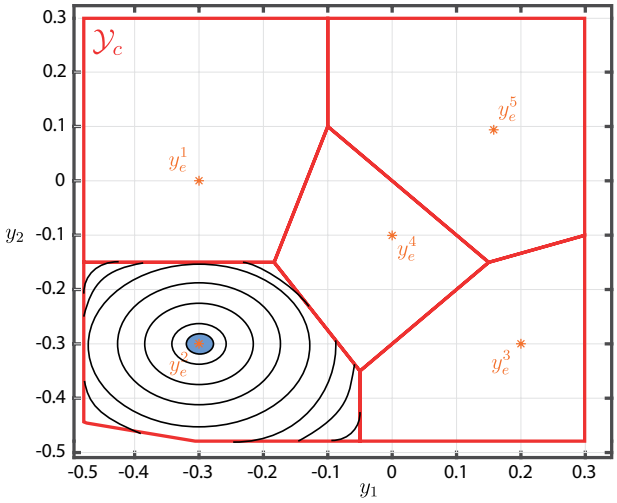


Figure 5.4: Family of robust one-step controllable sets covering a Voronoi partition

As a consequence, for definition of controllable sets, there exists a control law η_{sc} fulfilling the requirements (5.10). In particular, the safety control law $u^{sc}(k) = \eta_{sc}(y(k))$ can be obtained as follows:

- Given $y(k)$, find $\bar{l} \in \mathcal{L}$ such that $y(k) \in \mathcal{V}_{\bar{l}}$

- Determine the ROSC set containing $y(k)$ as

$$\bar{i} := \min_{0 \le i \le N_l} i : \; y(k) \in \mathcal{T}_i^{\bar{l}} \tag{5.16}$$

- **If** $\bar{i} == 0$, **then** $u^{sc}(k) = u^{\bar{l}}(k)$ as prescribed by (5.11)

- **Else** find $u^{sc}(k)$ solving the optimization problem

$$\begin{aligned} u^{sc}(k) = \arg\min_{u} \|Ay(k) + Bu - y_e^{\bar{l}}\|_2^2 \quad s.t. \\ Ay(k) + Bu \in \tilde{\mathcal{T}}_{\bar{i}-1}^{\bar{l}}, \; u \in \mathcal{U} \end{aligned} \tag{5.17}$$

Remark 5.2. *Please note that the above algorithm, by constructions, enjoys recursive feasibility of (5.17), see [79]. Moreover, most of the required computations (Voronoi partition* $\{\mathcal{V}_l\}_{l=1}^L$ *RCI sets* $\{\mathcal{T}_0^l\}_{l=1}^L$*, families of ROSC sets* $\{\mathcal{T}_i^l\}_{i=1}^{N_l}$*,* $l = 1, \ldots, L$*) can be carried into the offline phase, leaving online only the solution of a simple quadratic programming (QP) problem.* □

Proposition 5.1. *Consider the family of equilibrium pairs* $\{(x_e^l, u_e^l)\}_{l=1}^L$*, the Voronoi partition* $\{\mathcal{V}_l\}_{l=1}^L$*, the RCI sets* $\{\mathcal{T}_0^l\}_{l=1}^L$*, and the families of ROSC sets* $\{\mathcal{T}_i^l\}_{i=1}^{N_l}$*,* $l \in \mathcal{L}$ *computed according to (5.14)-(5.15). Then, if at* $k = k'$*, a persistent cyber-attack starts on the actuation channel and* $y(k') \in \mathcal{V}_l$*,* $1 \le l \le L$*, then the safety controller ensures that safety and recovery are guaranteed (see Definition 2.2 of chapter 2), and that for* $k \ge k' + N_l$ *the tracking error* $e(k) = Cx(k) - r(k)$ *is such that* $e(k) \le d^{sup}(C(\mathcal{T}_0^l \oplus (-\mathcal{M})), r(k)), \; \forall\, k \ge k' + N_l$*.*

Proof. Since $\bigcup_{l=1}^L \mathcal{V}_l = \mathcal{Y}_c$, and $\bigcup_{i=1}^{N_l} \mathcal{T}_i^l = \mathcal{V}_l$, $\forall\, l \in \mathcal{L}$, then $\forall\, y \in \mathcal{Y}_c, \exists\, l \in \mathcal{L}$ and $i \in \{0, \ldots, N_l\}$ such that $y \in \mathcal{T}_i^l$. As a consequence, the SC can be activated starting from any initial output condition. Moreover, by construction, if $y(k') \in \mathcal{V}_l$, then the SC will use the ROSC family $\{\mathcal{T}_i^l\}_{i=1}^{N_l}$ to determine admissible control actions $u^{sc}(k) \in \mathcal{U}$ as in (5.17) and ensure (given its recursive feasibility) that at each iteration, $y(k)$ moves in the successor of the current controllable set, until $y(k' + N_l) \in \mathcal{T}_0^l$. Therefore, $y(k)$ never leaves

the T-DoA domain $\mathcal{Y}_c \subseteq \mathcal{Y}$, allowing the networked controller to be safely re-enable at any time.

Moreover, if $y(k) \in \mathcal{T}_i^l$ is ROSC to $\mathcal{T}_{i-1}^l$, then, by considering the worst-case realization of the measurement noise $m(k) \in \mathcal{M}$, the corespondent state $x(k)$ is ROSC to $\mathcal{T}_{i-1}^l \oplus (-\mathcal{M})$ and $x(k'+N_l) \in \mathcal{T}_0^l \oplus (-\mathcal{M})$. Therefore, for $k \geq k' + N_l$, the system state is confined into the RPI region $\mathcal{T}_0^l \oplus (-\mathcal{M})$ and $y^r(k) = Cx(k)$ is confined into the set $C(\mathcal{T}_0^l \oplus (-\mathcal{M}))$. As a consequence, the maximum tracking error is equal to the maximum distance between the reference $r(k)$ and the RCI region confining $y^r(k)$, i.e., $C(\mathcal{T}_0^l \oplus (-\mathcal{M}))$, concluding the proof (see Figure. 5.5 as an example). □

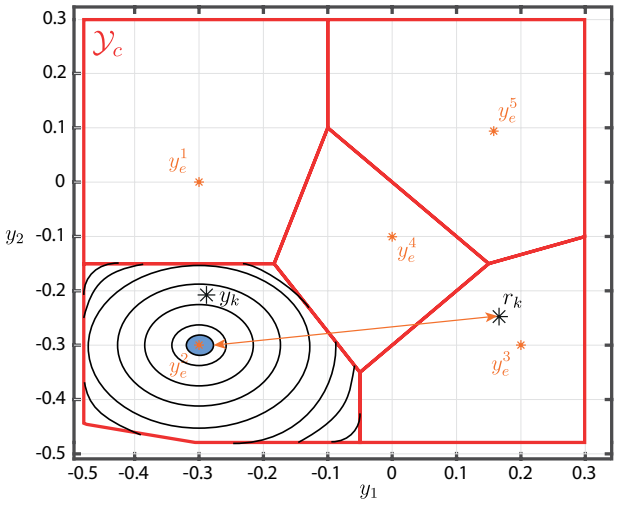


Figure 5.5: Maximum tracking error when the SC is activated

5.4.2 Tracking Supervisor

If a cyber-attack on the measurement channel starts at k', then $y(k'+k)$ is invalid for $0 \leq k \leq k_a$, with $k_a > 0$ the unknown attack duration. By exploiting the safety controller's capabilities, a simple option for the tracking supervisor would be to intentionally corrupt the integrity of the control signal $u(k')$, hence triggering the activation of the SC. Such a solution is effective in preserving the safety of the system and ensure recovery for $k > k_a$ (see Proposition 5.1). However, such an approach might not lead to the best outcome in

terms of tracking performance. For example, in Figure. 5.6, the above solution will force the plant to track the equilibrium output y_e^2 while perhaps other equilibrium points might be closer to the desired reference.

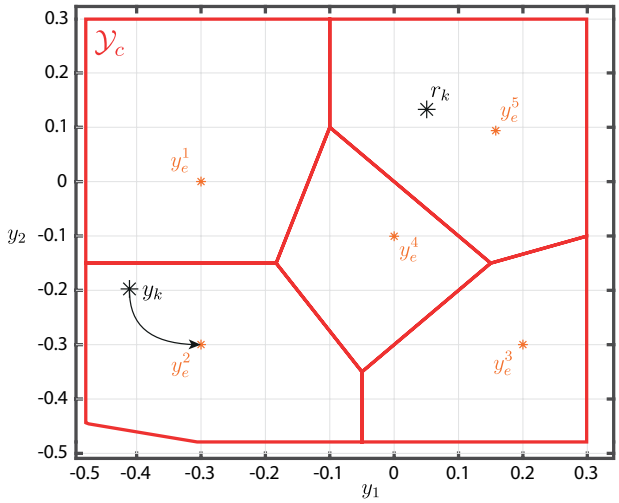


Figure 5.6: Output trajectory if the tracking supervisor intentionally corrupts the integrity of the control signal to activate the SC

The idea here pursued is that although $y(k'+k)$ is invalid for $0 \leq k \leq k_a$, its value can be estimated from $y(k'-1)$ (last valid measurement vector) by resorting to a worst-case approach. Such estimation can then be leveraged to allow, in a supervised fashion, the tracking controller to keep operating if there are the premises to achieve a better reference tracking (i.e., reaching a Voronoi partition closer to the current reference). Such an idea is here translated as follows.

First, we offline approximately quantify the tracking performance degradation associated with the safety controller actions. To this end, the following tracking index $I(i,j)$ is computed:

$$I(l_i, l_j) = \alpha I_1(l_i, l_j) + \beta I_2(l_i, l_j),\ \forall\,(l_i, l_j),\ l_i, l_j \in \mathcal{L} \tag{5.18}$$

where $\alpha \geq 0$ and $\beta \geq 0$ are two weighting factors (design parameters), and

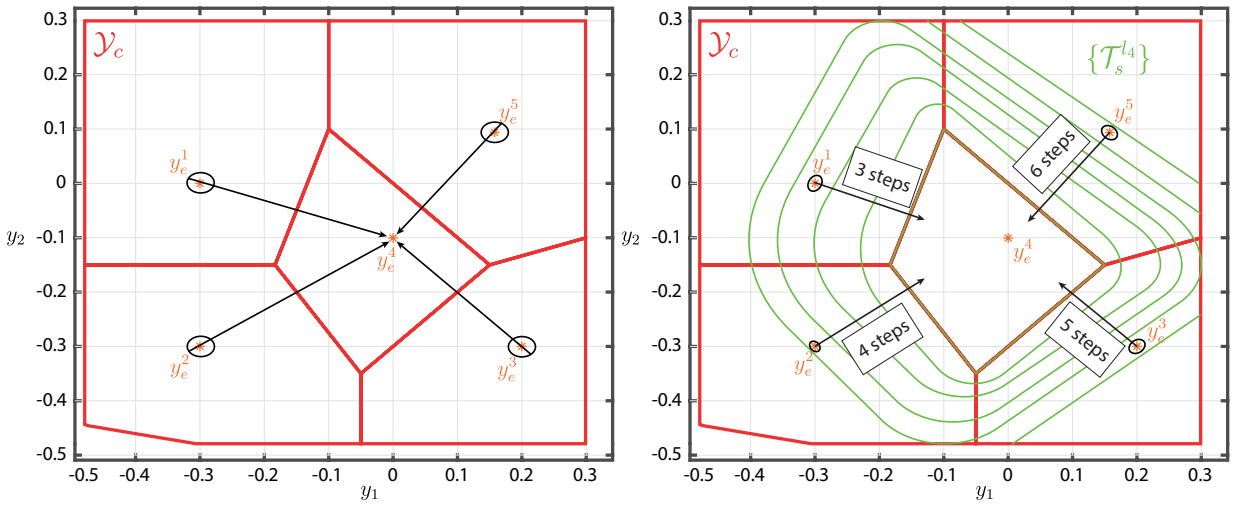
- $I_1(l_i, l_j) = d^{sup}(C(\mathcal{T}_0^{l_i} \oplus (-\mathcal{M}), Cx_e^{l_j})$. Such index quantifies the maximum tracking error if $y(k+k') \in \mathcal{T}_0^{l_i} \subseteq \mathcal{V}_{l_i}$ and $r(k)$ belongs to $\mathcal{V}_{l_j}$.

- $I_2(l_i, l_j) = \min_{0 \le p \le N_{l_j}} p : \mathcal{T}_0^{l_i} \subseteq \bigcup_{s=0}^{p} \{\mathcal{T}_s^{l_j}\}$, with $\{\mathcal{T}_s^{l_j}\}_{s=0}^{N_{l_j}}$ a set of $N_{l_j} \geq 0$ ROSC set built as

$$\mathcal{T}_s^{l_j} = \{y \in \mathcal{Y}_c : \exists u \in \mathcal{U} : Ay + Bu \in \tilde{\mathcal{T}}_{s-1}^{l_j}\}; \quad s \geq 1 \tag{5.19}$$

with starting RCI set $\mathcal{T}_0^{l_j} = \mathcal{V}_j$ and terminal condition $\mathcal{T}_0^{l_i} \subseteq \bigcup_{s=0}^{N_{l_j}} \{\mathcal{T}_s^{l_j}\}, \forall \mathcal{T}_0^{l_i}$. Such index quantifies the worst-case number of steps required for $y(k+k') \in \mathcal{T}_0^{l_i} \subseteq \mathcal{V}_{l_i}$ to enter the Voronoi partition $\mathcal{V}_{l_j}$ containing $\bar{y}_{r(k'+k)}$.

Remark 5.3. *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 $\|y^r(k) - r(k)\|$ committed activating the safety controller during the attack (see Proposition 5.1), while $I_2(l_i, l_j)$ approximates the time required to recover the reference tracking problem when the attack is terminated (see Figure. 5.7 as an example).* □



(a) $I_1(l_i, l_4); i = 1, 2, 3, 5$ (b) $I_2(l_i, l_4); i = 1, 2, 3, 5$

Figure 5.7: Graphical illustration of the meaning of $I_1(l_i, l_j)$ and $I_2(l_i, l_j)$ for a five region partition where $l_j = 4$ and $l_i = 1, \ldots, 4$

Let us denote with $\mathcal{V}_{l_y}$ and $\mathcal{V}_{l_r}$ the Voronoi sets containing $y(k'-1)$ and $\bar{y}_{r(k'-1)}$, respectively. 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,

$$\mathcal{I}(l_r) = [I(l_1, l_r), \ldots, I(l_y, l_r) \ldots, I(l_L, l_r)],$$
$$l_j \in \mathcal{L}, \forall\, j, \;\; I(l_1, l_r) \leq \ldots, \leq I(l_y, l_r) \leq \ldots, I(l_L, l_r)$$

Therefore, if $I(l_y, l_r) = I(l_1, l_r)$ then, the lowest tracking performance loss is obtained by forcing $y(k' + k)$ to remain in $\mathcal{V}_{l_y}$. On the other hand, if $I(l_y, l_r) \neq I(l_1, l_r)$, then better tracking performance are obtained if $y(k' + k)$, $k \geq 0$ can be steered into a pair (l_j, l_r) such that $I(l_j, l_r) < I(l_y, l_r)$.

While the first scenario ($I(l_y, l_r) = I(l_1, l_r)$) admits a simple solution, i.e., the invalidation of the integrity of $u(k')$ and the consequent activation of the emergency controller, the second one ($I(l_y, l_r) \neq I(l_1, l_r)$) is not trivial because:

- The measurement signal $y(k' + k)$ is invalid for $k \geq 0$
- An attack on the actuation channel could invalidated $u(k' + k)$ at any unpredictable time instant $k' + k$, $k \geq 0$

The first drawback implies that only a robust uncertain prediction of $y(k' + k)$ can be obtained from $y(k' - 1)$ for $k > 0$, while the second implies that to achieve the best tracking outcome it is not possible to rely on an optimization algorithm over a prediction horizon. As a consequence, here we resort to a worst-case approach where, given the uncertain predictions of $y(k' + k)$, the use or invalidation of the tracking controller action $u(k' + k)$ is decided given the chances (probability) that the robust one-step evolution produces a better tracking outcome.

Uncertain predictions: given $y(k' - 1)$, the output evolution of the system can be written (for linearity) as

$$y(k' + k) = \hat{y}(k' + k) + \tilde{y}(k' + k), \quad k \geq 0 \tag{5.20}$$

with

$$\hat{y}(k'+k) = A^{k+1}y(k'-1) + \sum_{j=0}^{k}\left(A^{j}Bu(k'+k-1-j)\right) \quad (5.21)$$

$$\tilde{y}(k'+k) = \sum_{j=0}^{k}\left(A^{j}B_{d}d(k'+k-1-j)\right) - A^{k+1}m(k'-1) + m(k'+k) \quad (5.22)$$

and where $\hat{y}(k'+k)$ denotes the predictable output evolution due to y and u, while $\tilde{y}(k'+k)$ represents the evolution due to d and m. Therefore, we have that

$$y(k'+k) \in \hat{y}(k'+k) \oplus \mathcal{N}(k'+k),\ \ \mathcal{N}(k'+k) = \sum_{j=0}^{k}\left(A^{j}B_{d}\mathcal{D}\right) \oplus (-A^{k+1}\mathcal{M}) \oplus \mathcal{M} \quad (5.23)$$

where $\mathcal{N}(k'+k)$ defines an the uncertainty set about the estimated output at $k'+k$.

Tracking performance evaluation: Given the disturbance-free prediction $\hat{y}(k'+k)$, and $r(k'+k)$, the tracking controller (5.7) computes the control input $u(k'+k)$. Such command is then evaluated in terms of associated tracking performance. Please note that the uncertainty $\mathcal{N}(k'+k)$ does not allow us to evaluate, in a deterministic (single vector) manner, the tracking index $I(l_i, l_j)$. Therefore, a modified weighted index $J(k'+k+1)$ will be here used. In particular, two information are computed:

$$\mathcal{L}^{w}(k'+k+1) := \{l \in \mathcal{L} :\ I(l, l_{r(k'+k)}) > I(l_{y(k'+k+1)}, l_{r(k'+k)}), \quad (5.24)$$
$$(\hat{y}(k'+k+1) \oplus \mathcal{N}(k'+k+1)) \cap \mathcal{V}_{l} \neq \emptyset\}$$

and

$$J(k'+k+1) = \sum_{l_j \in \mathcal{L}\setminus\mathcal{L}^{w}(k+1)} \frac{vol\left((\hat{y}(k'+k+1) \oplus \mathcal{N}(k'+k+1)) \cap \mathcal{V}_{j}\right)}{vol\left(\mathcal{N}(k'+k+1)\right)} I(l_j, l_{r(k'+k)}) \quad (5.25)$$

with $vol(\cdot)$ a function computing the volume. Please note that $\mathcal{L}^{w}(k'+k+1)$ defines the set of Voronoi regions, with a tracking performance index I higher (worse) than the current one, intersected by the robust one-step prediction (see Figure. 5.8 as an example).

Moreover, $J(k'+k+1)$ defines a weighted sum of the tracking index according to the volume overlap between the uncertain prediction set and the Voronoi regions.

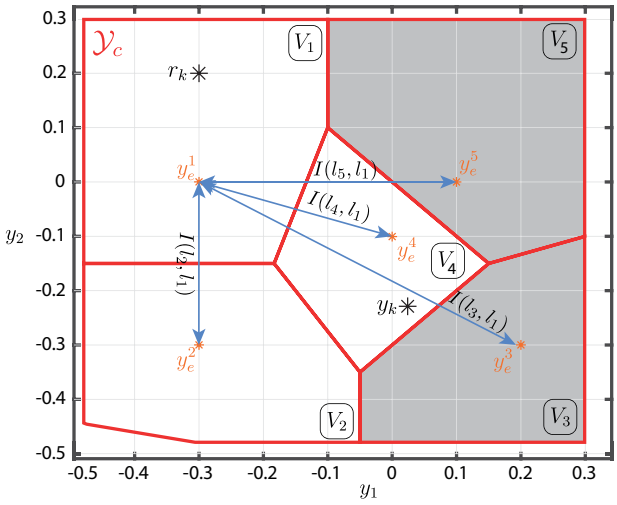


Figure 5.8: Voronoi partitions and tracking performance index w.r.t. the current reference signal r_k. Gray regions are the ones with a worse tracking index w.r.t to V_4 (i.e., the region containing y_k).

Proposition 5.2. *Consider the current reference signal $r(k+k')$, the predicted uncertain measurement $y(k'+k)$ in (5.23), the set $\mathcal{L}^w(k'+k+1)$ in (5.24) and the cost $J(k'+k+1)$ in (5.25). If $(\hat{y}(k'+k)\oplus\mathcal{N}(k'+k))\subseteq\mathcal{Y}_c$ and $\mathcal{L}^w(k'+k+1)=\emptyset$, then $u(k'+k)$ ensures that, in the worst-case scenario, the tracking performance is not worse (according to (5.18)) than the one obtained using $u^{sc}(k'+k)$ (e.g., activating SC). Moreover, if, in addition, $J(k'+k+1)<J(k'+k)$, then $u(k'+k)$ increases the chances that $y(k'+k+1)$ will be inside a Voronoi region with better tracking performance (according to (5.18)) than the current one.*

Proof. If $\mathcal{L}^w(k'+k+1)\neq\emptyset$, then there is a possibility that the one-step output evolution will be in a Voronoi region with tracking index I higher (worse) than the current one, i.e., there exist an admissible disturbance realization $d\in\mathcal{D}$, $m\in\mathcal{M}$, $y(k'+k+1)\in\mathcal{V}_l$ with l such that $I(l,l_{r(k'+k)})\geq I(l_{y(k'+k+1)},l_{r(k'+k)})$. Moreover, if $(\hat{y}(k'+k+1)\oplus\mathcal{N}(k'+k+1))\not\subseteq\mathcal{Y}_c$, then, there exist an admissible disturbance realization $d\in\mathcal{D}$, $m\in\mathcal{M}$ such that the one-step evolution goes outside of the T-DoA domain $\mathcal{Y}_c$. As a consequence, the tracking

performance degradation in the worst-case scenario are lower if $y(k'+k+1)$ is confined in $\mathcal{V}_{l_{y(k'+k)}}$, i.e., if the safety controller (5.9) is activated. On the other hand, if $\mathcal{L}^w(k'+k+1) = \emptyset$, $(\hat{y}(k'+k+1) \oplus \mathcal{N}(k'+k+1)) \subseteq \mathcal{Y}_c$, and $J(k'+k+1) < J(k'+k)$, then by applying $u(k'+k)$ there is no possible that $y(k'+k+1)$ will be a region $\mathcal{V}_l$ with a worse tracking outcome or go outside of the T-DoA. Moreover, since $J(k'+k+1)$ computes a weighted sum of the tracking cost $I(l_j, l_{r(k'+k)})$ by means of the volume percentage of the uncertain output set $(\hat{y}(k'+k+1) \oplus \mathcal{N}(k'+k+1))$ intersecting the Voronoi cell $\mathcal{V}_j$, we can conclude that if $J(k'+k+1) < J(k'+k)$ then by applying $u(k'+k)$, we have, in the worst-case scenario, higher probabilities of achieving an improved tracking performance (5.18). □

Given the results in Proposition 5.2, under a cyber-attack on the measurement channel starting at k', the tracking supervisor logic can be summarized as follows:

1. The predicted output $\hat{y}(k'+k)$ is obtained as in (5.21) and used by the tracking controller (5.7) to obtain $u(k'+k)$;

2. **If** $(\hat{y}(k'+k+1) \oplus \mathcal{N}(k'+k+1)) \not\subseteq \mathcal{Y}_c$, or ($\mathcal{L}(k'+k+1)^w \neq \emptyset$) or ($\mathcal{L}(k'+k+1)^w = \emptyset$ and $J(k'+k+1) > J(k'+k)$) **then** the integrity of $u(k'+k)$ is intentionally compromised and the safety controller (5.9) activated;

3. **Else** $u(k'+k)$ is sent over the actuation channel.

5.4.3 Implementation is the absence of MAC

Although outside of the scope of this technical note, in this subsection, we discuss (for the sake of completeness) how it is possible to adapt the proposed solution to the case where MAC is not available, and a control-theoretical anomaly detector module [34] is used to detect the presence of attacks. For the following, we generically model the detection mechanisms as capable of providing attack detection with a bounded delay $0 \leq \bar{\tau} < \infty$. Please note that if the detection mechanism is such that $\bar{\tau} = 0$, then its attack detection capability is equivalent to using MAC.

If $\bar{\tau} > 0$, then the proposed safety controller and tracking supervisor actions must be properly modified to ensure their correct operations and the safety of the system:

Safety Controller. This module is responsible for the plant's safety. Therefore, it must be equipped with a local attack detector capable of revealing an attack before the state trajectory leaves the T-DoA $\mathcal{Y}_c$. To this end, a robust safety risk detection rule can be obtained exploiting the concept of robust one-step reachable set. In particular, when $u(k)$ is received, then the following rule can be used to decide if to trust $u(k)$ or to activate the SC.

$$\textbf{if } \mathcal{Y}^+ \subseteq \mathcal{Y}_c \textbf{ then } \text{apply } u(k) \textbf{ else } \text{ activate SC} \tag{5.26}$$

with $\mathcal{Y}^+ := Ay(k) + Bu(k) \oplus B_d\mathcal{D} \oplus (-A)\mathcal{M} \oplus \mathcal{M}$.

Tracking Supervisor. This module is in charge of supervising the worst-case open-loop evolution of the plant starting from the last valid measurement. If $\bar{\tau} > 0$, then the last trusted measurement is $y(k' - 1 - \bar{\tau})$. Therefore, to minimize the cost index (5.25), the trajectory predictions (5.21)-(5.22) must be replaced with the followings:

$$\hat{y}(k'+k) = A^{k+1+\bar{\tau}} y(k'-1-\bar{\tau}) + \sum_{j=0}^{k+\bar{\tau}} \big(A^j B u(k'+k-1-j)\big) \tag{5.27}$$

$$\tilde{y}(k'+k) = \sum_{j=0}^{k+\bar{\tau}} \big(A^j B_d d(k'+k-1-j)\big) + A^{k+1+\bar{\tau}} m_(k'-1-\bar{\tau}) + m(k'+k) \tag{5.28}$$

and the uncertainty associated with the predictions is

$$\mathcal{N}_\tau(k'+k) = \sum_{j=0}^{k+\bar{\tau}} \big(A^j B_d \mathcal{D}\big) \oplus (-A^{k+1+\bar{\tau}}\mathcal{M}) \oplus \mathcal{M} > \mathcal{N}(k'+k) \tag{5.29}$$

Please note that the uncertainty $\mathcal{N}_\tau(k'+k)$ increases with the delay $\bar{\tau}$. Consequently, the conservativeness of the tracking supervisor actions increases with $\bar{\tau}$.

5.5 Simulation Example

In this section, we consider as testbed for the proposed approach the Two-Tank water system used in [110]. The states of the system are given by the water's levels in the two tanks i.e., $x = [h_1, h_2]^T$, and the control input vector $u = [u_p, u_l, u_u]^T$ consists of the pump (u_p) and valves (u_l, and u_u) signals. The nonlinear continuous-time dynamics have been linearized around the equilibrium pair $x_{eq} = [0.5, 0.5]^T$, and $u_{eq} = [0.938, 1, 0.833]^T$ and discretized with a sampling time $T_s = 1sec$. The obtained model (5.1) has the following matrices

$$A = \begin{bmatrix} 0.993 & 0.003 \\ 0.007 & 0.982 \end{bmatrix}, B = \begin{bmatrix} 0.008 & -0.003 & -0.003 \\ 0 & 0.003 & 0.003 \end{bmatrix} \tag{5.30}$$
$$B_p = I_2,\ C = I_2$$

The process and measurement disturbance sets are $\mathcal{M} = \mathcal{D} = \{d \in \mathbb{R}^2 : |d(j)| \leq 0.001,\ j = 1, 2\}$, while the state and input constraints are $-0.5 \leq u_p \leq 1.5$, $-0.25 \leq u_l \leq 1.75$, $-0.8 \leq u_u \leq 1.2$, $0.02 \leq h_1, h_2 \leq 0.8$. The tracking controller (5.7) is a Command Governor (CG) [46] and its T-DoA $\mathcal{Y}_c$ is shown in Figure. 5.9. The safety controller is configured with a five region Voronoi partition of $\mathcal{Y}_c$ (see Figure. 5.9) obtained using as generators the equilibrium states $y_e^1 = [-0.3, 0]^T$, $y_e^2 = [-0.3, -0.3]^T$, $y_e^3 = [0.2, -0.3]^T$, $y_e^4 = [0, -0.1]^T$, $y_e^5 = [0.1, 0]^T$. On the other hand, the tracking supervisor is configured to use $\alpha = 1$, $\beta = 0$ in (5.18).

In the performed simulations, the plant is required to track a time-varying reference signal $r(k)$ (see Figure. 5.10) while three attacks on the measurement channel occur. The first attack for $80 \leq t < 230\,\text{sec}$ constantly invalidate the measurement vector $y(k)$. Since at $t = 80\,\text{sec}$, $y(k) \in \mathcal{V}_3$ and $r(k) \in \mathcal{V}_5$, the tracking supervisor is activated. Therefore, according to its logic, at each time instant, the predicted measurement output $\hat{y}(80 + k)$ is used, the performance index $J(80 + k + 1)$ evaluated and the output estimation error $\mathcal{N}(80 + k + 1)$ computed (green regions in Figure. 5.9). It is possible to notice that along

the output trajectory the index $J(80+k+1)$ has a monotonically non-increasing behavior. In particular, it remains constantly equals to $J(80) = 0.3299$ until the uncertain output set $\hat{y}(80+k+1) \oplus \mathcal{N}(80+k+1)$ intersects $\mathcal{V}_5$. After that, since $I_1(5,5) < I_1(3,5)$, the index J starts decreasing denoting that the tracking performance are improving. The tracking supervisor actions proceed until $t = 155$, where $\hat{y}(156) \oplus \mathcal{N}(156) \nsubseteq \mathcal{Y}_c$. In this scenario, for safety reasons, the control input is invalidated and the safety controller activated. As a consequence for $155 \leq t < 230$, the safety controller confines the output trajectory in the RPI region $\mathcal{T}_0^5$ centered in y_e^5 and the reference is not tracked. However, when the first attacks ends ($t = 229\,\text{sec}$), the plant is capable of recovering its tracking task (see Figures 5.9-5.10). A similar reasoning applies to the second attack for $300 \leq t < 380$. The difference in this case is that the uncertain output set never violates the safety constraints (pink sets in Figure. 5.9). As a consequence, the tracking task is never suspended. The third attack intermittently affects the measurement channel for $500 \leq t < 600$. In this case, given the nature of the attack, the sporadically received measures allows the tracking supervisor to reset the uncertainty set (yellow sets in Figure. 5.9), so avoiding the suspension of the tracking task. Finally, in Figure. 5.10 and Table 5.1 the proposed solution is contrasted with the one in [67]. Since in [67] 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 bigger tracking performance loss. By measuring the tracking error e_r as $e_r = \sum_{k=1}^{N_s} \frac{\|y^r(k)-r(k)\|}{N_s}$, with N_s the simulation steps, Table 5.1 reports the obtained numerical results confirming that, compared to [67], the proposed solution reduces the tracking performance degradation.

Table 5.1: Tracking error: proposed approach, [67], no attack

	No attack	Proposed Approach	[67]
e_r	0.0717	0.0893	0.1662

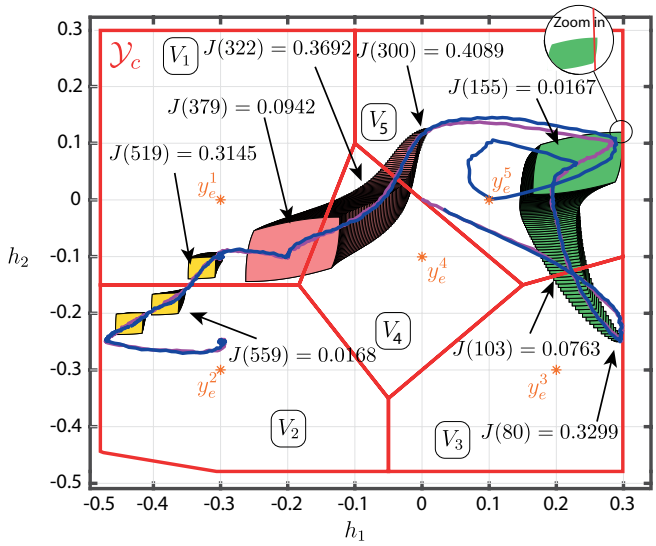


Figure 5.9: Output trajectory: proposed solution with attacks (blue solid line) vs trajectory in attack-free scenario (purple solid line).

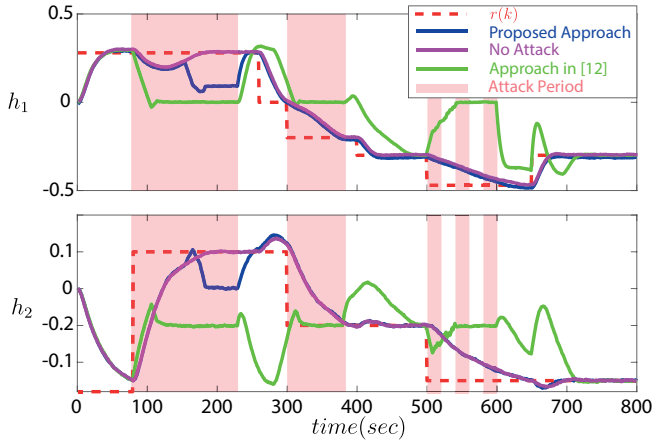


Figure 5.10: State evolution: no attack, proposed approach, [67].

5.6 Conclusion

In this chapter, 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, one local to the tracking controller) whose aim is to preserve safety while improving, in a supervised fashion, the tracking performance under attacks. The obtained theoretical and simulation results have shown the features of the proposed scheme.

Chapter 6

Conclusion and Future Works

In this thesis, first, we discussed three different research works in the field of safety and security of CPSs , i.e., we have shown the existence of finite-time covert attacks, designed a safety preserving control architecture, and faced the reference tracking problem under cyber-attacks.

In chapter 3, we designed a new type of attacks, namely finite-time covert attacks, targeting constrained and unconstrained control systems. The peculiar capability of such a class of attacks is that they remain stealthy also in the post-attack phase. We shown that if unconstrained control systems are considered, then such an attack can be designed resorting to a three-phase covert attack by properly leveraging a FIR state estimator and reachability arguments. In the constrained case, such an attack has been designed by jointly combining robust controllability arguments, and a set-theoretic-based receding horizon control paradigm.

In chapter 4, we proposed a novel networked control architecture in order to ensure plant safety under different cyber-attacks affecting the communication channels. This architecture has been designed by taking advantage of (i) an anomaly detector, local to the networked controller, capable of revealing FDI attacks on the control input, sensor measurement, and setpoint signals, and (ii) a smart actuator, local to the plant, capable of activating an ad-hoc designed emergency controller at least one-step before any plant

safety constraint could be violated. It has been formally proved that plant safety is achieved and guaranteed in the proposed architecture, regardless of the attacker's actions and detector performance.

In chapter 5, a robust solution to the safety and reference tracking for constrained CPSs under attacks has been proposed. The obtained solution is capable of minimizing the tracking performance degradation under attacks while preserving the plant's safety. The latter is achieved by leveraging robust reachability arguments and a Voronoi partition of the tracking domain. The peculiar capability of such a solution is that it consists of two add-on modules (one local to the plant, one local to the tracking controller) that can be potentially installed in any control system for CPS.

6.1 Future research directions

Some possible future research directions in which the research, done in this thesis, can be extended and improved are outlined below:

- The control techniques developed in Chapter 3 to show the existence of finite-time covert attacks can be used to prove the existence of other classes of finite-time undetectable attacks.

- The safety preserving architecture proposed Chapter 4 and 5 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 performance of the theoretical results in this thesis can be validated on real CPSs.

Bibliography

[1] Y. Mo, T.-J. Kim, B. Kenneth, D. Dickinson, H. Lee, A. Perrig, and B. Sinopoli, "Cyber–physical security of a smart grid infrastructure," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 195–209, 2012.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] V. 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.

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

[10] 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.

[11] 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.

[12] Y.-L. Huang, A. A. Cárdenas, S. Amin, Z.-S. Lin, H.-Y. Tsai, and S. Sastry, "Understanding the physical and economic consequences of attacks on control systems," *International Journal of Critical Infrastructure Protection*, vol. 2, no. 3, pp. 73–83, 2009.

[13] "Compromise of u.s. water treatment facility," *cybersecurity and infrastructure security agency*, 2021. [Online]. Available: https://us-cert.cisa.gov/ncas/alerts/aa21-042a

[14] A. Hassanzadeh, A. Rasekh, S. Galelli, M. Aghashahi, R. Taormina, A. Ostfeld, and M. K. Banks, "A review of cybersecurity incidents in the water sector," *Journal of Environmental Engineering*, vol. 146, no. 5, 2020.

[15] J. Slay and M. Miller, "Lessons learned from the maroochy water breach," in *International conference on critical infrastructure protection*. Springer, 2007, pp. 73–82.

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

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

[18] Y. Shoukry, P. Martin, P. Tabuada, and M. Srivastava, "Non-invasive spoofing attacks for anti-lock braking systems," in *International Workshop on Cryptographic Hardware and Embedded Systems*. Springer, 2013, pp. 55–72.

[19] 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.

[20] A. Teixeira, D. Pérez, H. Sandberg, and K. Johansson, "Attack models and scenarios for networked control systems," in *International conference on High Confidence Networked Systems*. ACM, 2012, pp. 55–64.

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

[22] S. Amin, X. Litrico, S. Sastry, and A. M. Bayen, "Cyber security of water scada systems—part i: Analysis and experimentation of stealthy deception attacks," *IEEE Transactions on Control Systems Technology*, vol. 21, no. 5, pp. 1963–1970, 2012.

[23] G. Park, C. Lee, H. Shim, Y. Eun, and K. H. Johansson, "Stealthy adversaries against uncertain cyber-physical systems: Threat of robust zero-dynamics attack," *IEEE Transactions on Automatic Control*, vol. 64, no. 12, pp. 4907–4919, 2019.

[24] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "Revealing stealthy attacks in control systems," in *Annual Allerton Conference on Communication, Control, and Computing*. IEEE, 2012, pp. 1806–1813.

[25] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *Annual Allerton Conference on Communication, Control, and Computing*. IEEE, 2009, pp. 911–918.

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

[27] D. Mikhaylenko and P. Zhang, "Stealthy local covert attacks on cyber-physical systems," in *American Control Conference (ACC)*, 2020, pp. 2568–2573.

[28] A. Abdelwahab, W. Lucia, and A. Youssef, "Covert channels in cyber-physical systems," *IEEE Control Systems Letters*, vol. 5, no. 4, pp. 1273–1278, 2020.

[29] H. Jeon and Y. Eun, "A stealthy sensor attack for uncertain cyber-physical systems," *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 6345–6352, 2019.

[30] K. Gheitasi and W. Lucia, "A finite-time stealthy covert attack against cyber-physical systems," in *International Conference on Control, Decision and Information Technologies (CoDIT)*, vol. 1. IEEE, 2020, pp. 347–352.

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

[32] 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.

[33] A. Cardenas, S. Amin, B. Sinopoli, A. Giani, A. Perrig, S. Sastry *et al.*, "Challenges for securing cyber physical systems," in *Workshop on future directions in cyber-physical systems security*, vol. 5, no. 1. Citeseer, 2009.

[34] J. Giraldo, D. Urbina, A. Cardenas, J. Valente, M. Faisal, J. Ruths, N. O. Tippenhauer, H. Sandberg, and R. Candell, "A survey of physics-based attack detection in cyber-physical systems," *ACM Computing Surveys*, vol. 51, no. 4, pp. 1–36, 2018.

[35] 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.

[36] 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)*, 2017, pp. 1374–1379.

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

[38] F. Miao, Q. Zhu, M. Pajic, and G. 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, 2017.

[39] 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.

[40] R. Romagnoli, S. Weerakkody, and B. Sinopoli, "A model inversion based watermark for replay attack detection with output tracking," in *American Control Conference (ACC)*, 2019, pp. 384–390.

[41] N. R. Chowdhury, J. Belikov, D. Baimel, and Y. Levron, "Observer-based detection and identification of sensor attacks in networked cpss," *Automatica*, vol. 121, 2020.

[42] J. Zhang, L. Pan, Q.-L. Han, C. Chen, S. Wen, and Y. Xiang, "Deep learning based attack detection for cyber-physical system cybersecurity: A survey," *IEEE/CAA Journal of Automatica Sinica*, 2021.

[43] X. Zhang, Y. Lu, and M. Zhu, "Attack detection of nonlinear distributed control systems," in *American Control Conference (ACC)*. IEEE, 2020, pp. 1459–1464.

[44] W. Lucia, K. Gheitasi, and M. Ghaderi, "Setpoint attack detection in cyber-physical systems," *IEEE Transactions on Automatic Control*, 2020.

[45] ——, "A command governor based approach for detection of setpoint attacks in constrained cyber-physical systems," in *IEEE Conference on Decision and Control (CDC)*, 2018, pp. 4529–4534.

[46] A. Bemporad, A. Casavola, and E. Mosca, "Nonlinear control of constrained linear systems via predictive reference management," *IEEE Transactions on Automatic Control*, vol. 42, no. 3, pp. 340–349, 1997.

[47] A. Bemporad, "Reference governor for constrained nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 43, no. 3, pp. 415–419, 1998.

[48] E. G. Gilbert, I. Kolmanovsky, and K. T. Tan, "Discrete-time reference governors and the nonlinear control of systems with state and control constraints," *International Journal of Robust and Nonlinear Control*, vol. 5, no. 5, pp. 487–504, 1995.

[49] 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.

[50] L. An and G.-H. Yang, "Secure state estimation against sparse sensor attacks with adaptive switching mechanism," *IEEE Transactions on Automatic Control*, vol. 63, no. 8, pp. 2596–2603, 2017.

[51] S. Zonouz, K. Rogers, R. Berthier, R. Bobba, W. Sanders, and T. Overbye, "Scpse: Security-oriented cyber-physical state estimation for power grid critical infrastructures," *IEEE Transactions on Smart Grid*, vol. 3, no. 4, pp. 1790–1799, 2012.

[52] 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.

[53] Y. Gao, G. Sun, J. Liu, Y. Shi, and L. Wu, "State estimation and self-triggered control of cpss against joint sensor and actuator attacks," *Automatica*, vol. 113, p. 108687, 2020.

[54] W. Heemels, A. Teel, N. Van de Wouw, and D. Nesic, "Networked control systems with communication constraints: Tradeoffs between transmission intervals, delays and performance," *IEEE Transactions on Automatic control*, vol. 55, no. 8, pp. 1781–1796, 2010.

[55] 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.

[56] T. Li, B. Chen, L. Yu, and W.-A. Zhang, "Active security control approach against dos attacks in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 66, no. 9, pp. 4303–4310, 2020.

[57] 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.

[58] S. Feng, A. Cetinkaya, H. Ishii, P. Tesi, and C. De Persis, "Networked control under dos attacks: Tradeoffs between resilience and data rate," *IEEE Transactions on Automatic Control*, vol. 66, no. 1, pp. 460–467, 2020.

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

[60] 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.

[61] W. Lucia, B. Sinopoli, and G. Franze, "A set-theoretic approach for secure and resilient control of cyber-physical systems subject to false data injection attacks," in *Science of Security for Cyber-Physical Systems Workshop*, 2016, pp. 1–5.

[62] J. Milošević, A. Teixeira, K. H. Johansson, and H. Sandberg, "Actuator security indices based on perfect undetectability: Computation, robustness, and sensor placement," *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3816–3831, 2020.

[63] L. Zhai and K. G. Vamvoudakis, "A data-based private learning framework for enhanced security against replay attacks in cyber-physical systems," *International Journal of Robust and Nonlinear Control*, vol. 31, no. 6, pp. 1817–1833, 2021.

[64] Y. Joo, Z. Qu, and T. Namerikawa, "Resilient control of cyber-physical system using nonlinear encoding signal against system integrity attacks," *IEEE Transactions on Automatic Control*, vol. 66, no. 9, pp. 4334–4341, 2020.

[65] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, "Malicious data attacks on the smart grid," *IEEE Transactions on Smart Grid*, vol. 2, no. 4, pp. 645–658, 2011.

[66] K. Gheitasi and W. Lucia, "Undetectable finite-time covert attack on constrained cyber-physical systems," *IEEE Transactions on Control of Network Systems*, 2022.

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

[68] 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)*. IEEE, 2019, pp. 1449–1454.

[69] M. Ghaderi, K. Gheitasi, and W. Lucia, "A novel control architecture for the detection of false data injection attacks in networked control systems," in *American Control Conference (ACC)*. IEEE, 2019, pp. 139–144.

[70] F. Blanchini and S. Miani, "Any domain of attraction for a linear constrained system is a tracking domain of attraction," *SIAM Journal on Control and Optimization*, vol. 38, no. 3, pp. 971–994, 2000.

[71] O. Katsuhiko, *Modern control engineering*. Prentice Hall, 2010.

[72] R. Tunga, C. Murguia, and J. Ruths, "Tuning windowed chi-squared detectors for sensor attacks," in *American Control Conference (ACC)*. IEEE, 2018, pp. 1752–1757.

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

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

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

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

[77] S. Weerakkody, Y. Mo, and B. Sinopoli, "Detecting integrity attacks on control systems using robust physical watermarking," in *IEEE Conference on Decision and Control (CDC)*, 2014, pp. 3757–3764.

[78] F. Miao, Q. Zhu, M. Pajic, and G. Pappas, "Coding sensor outputs for injection attacks detection," in *IEEE Conference on Decision and Control (CDC)*, 2014, pp. 5776–5781.

[79] 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.

[80] M. Herceg, M. Kvasnica, C. Jones, and M. Morari, "Multi-Parametric Toolbox 3.0," in *Proc. of the European Control Conference*, Zürich, Switzerland, July 17–19 2013, pp. 502–510, http://control.ee.ethz.ch/~mpt.

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

[82] S.-C. Huang, Y.-L. Lo, and C.-N. Lu, "Non-technical loss detection using state estimation and analysis of variance," *IEEE Transactions on Power Systems*, vol. 28, no. 3, pp. 2959–2966, 2013.

[83] W. H. Kwon, P. Kim, and S. H. Han, "A receding horizon unbiased fir filter for discrete-time state space models," *Automatica*, vol. 38, no. 3, pp. 545–551, 2002.

[84] W. H. Kwon, P. S. Kim, and P. Park, "A receding horizon kalman fir filter for discrete time-invariant systems," *IEEE Transactions on Automatic Control*, vol. 44, no. 9, pp. 1787–1791, 1999.

[85] 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.

[86] R. S. Smith, "A decoupled feedback structure for covertly appropriating networked control systems," *IFAC Proceedings Volumes*, vol. 44, no. 1, pp. 90–95, 2011.

[87] W. Lucia, D. Famularo, and G. Franze, "A set-theoretic reconfiguration feedback control scheme against simultaneous stuck actuators," *IEEE Transactions on Automatic Control*, vol. 63, no. 8, pp. 2558–2565, 2017.

[88] J. K. Scott, D. M. Raimondo, G. R. Marseglia, and R. D. Braatz, "Constrained zonotopes: A new tool for set-based estimation and fault detection," *Automatica*, vol. 69, pp. 126–136, 2016.

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

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

[91] A. Kurzhanskiĭ and I. Vályi, *Ellipsoidal calculus for estimation and control*. Nelson Thornes, 1997.

[92] A. H. Glattfelder and W. Schaufelberger, *Control systems with input and output constraints*. Springer Science & Business Media, 2003.

[93] T. Nguyen and F. Jabbari, "Disturbance attenuation for systems with input saturation: an lmi approach," *IEEE Transactions on Automatic Control*, vol. 44, no. 4, pp. 852–857, 1999.

[94] G. Franze, W. Lucia, and F. Tedesco, "Resilient model predictive control for constrained cyber-physical systems subject to severe attacks on the communication

channels," *IEEE Transactions on Automatic Control*, 2021. [Online]. Available: 10.1109/TAC.2021.3084237

[95] H. Gao, T. Chen, and L. Wang, "Robust fault detection with missing measurements," *International Journal of Control*, vol. 81, no. 5, pp. 804–819, 2008.

[96] 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.

[97] S. V. Rakovic and M. Baric, "Parameterized robust control invariant sets for linear systems: Theoretical advances and computational remarks," *IEEE Transactions on Automatic Control*, vol. 55, no. 7, pp. 1599–1614, 2010.

[98] X. Jin, W. M. Haddad, and T. Yucelen, "An adaptive control architecture for mitigating sensor and actuator attacks in cyber-physical systems," *IEEE Transaction on Automatic Control*, vol. 62, no. 11, pp. 6058–6064, 2017.

[99] G. Franzè, F. Tedesco, and W. Lucia, "Resilient control for cyber-physical systems subject to replay attacks," *IEEE Control Systems Letters*, vol. 3, no. 4, pp. 984–989, 2019.

[100] Z. Artstein and S. V. Raković, "Feedback and invariance under uncertainty via set-iterates," *Automatica*, vol. 44, no. 2, pp. 520–525, 2008.

[101] D. Liberzon and A. S. Morse, "Basic problems in stability and design of switched systems," *IEEE control systems magazine*, vol. 19, no. 5, pp. 59–70, 1999.

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

[103] S. M. Dibaji, M. Pirani, D. B. Flamholz, A. M. Annaswamy, K. H. Johansson, and A. Chakrabortty, "A systems and control perspective of cps security," *Annual Reviews in Control*, 2019.

[104] S. Koteshwara, A. Das, and K. K. Parhi, "Fpga implementation and comparison of aes-gcm and deoxys authenticated encryption schemes," in *IEEE International Symposium on Circuits and Systems*. IEEE, 2017, pp. 1–4.

[105] A. M. Naseri, W. Lucia, M. Mannan, and A. Youssef, "On securing cloud-hosted cyber-physical systems using trusted execution environments," in *IEEE International Conference on Autonomous Systems (ICAS)*, 2021, pp. 1–5. [Online]. Available: 10.1109/ICAS49788.2021.9551155

[106] C. Murguia, F. Farokhi, and I. Shames, "Secure and private implementation of dynamic controllers using semihomomorphic encryption," *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3950–3957, 2020.

[107] J. Tran, F. Farokhi, M. Cantoni, and I. Shames, "Implementing homomorphic encryption based secure feedback control," *Control Engineering Practice*, vol. 97, 2020.

[108] M. S. Darup, A. Redder, I. Shames, F. Farokhi, and D. Quevedo, "Towards encrypted mpc for linear constrained systems," *IEEE Control Systems Letters*, vol. 2, no. 2, pp. 195–200, 2017.

[109] M. S. Darup, A. B. Alexandru, D. E. Quevedo, and G. J. Pappas, "Encrypted control for networked systems: An illustrative introduction and current challenges," *IEEE Control Systems Magazine*, vol. 41, no. 3, pp. 58–78, 2021.

[110] J. H. Richter, W. Heemels, N. van de Wouw, and J. Lunze, "Reconfigurable control of piecewise affine systems with actuator and sensor faults: stability and tracking," *Automatica*, vol. 47, no. 4, pp. 678–691, 2011.