

Imperial College of Science, Technology and Medicine

**Plug-and-play
fault-tolerant and cyber-secure control:
application to future distribution networks**

Alexander J. Gallo

Supervisors: Prof T. Parisini, Dr F. Boem

Submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy at Imperial College of London

Control and Power Research Group
Department of Electrical and Electronic Engineering
Imperial College of London
September 2020

Declaration of originality and copyright statement

I hereby declare that this thesis is the fruit of my own work. Any idea, concept or result presented here that is the outcome of work that is not my own is appropriately referenced, be it published or otherwise.

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Licence (CC BY-NC-SA).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author, do not use it for commercial purposes and share any derivative works under the same licence.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Alexander J. Gallo

Abstract

In this thesis, we study the development of distributed cyber-attack detection methodologies with plug-and-play characteristics, to monitor large-scale cyber-physical systems, with a particular focus on future energy distribution networks.

Over the last few decades, the penetration of distributed generation capacity in power networks has greatly increased, especially that of renewable energy sources. Thus, to maintain the required performance, novel methods for voltage and power regulation have been suggested.

Given their scale and complexity, distributed control architectures have been proposed to regulate these networks, where the system is divided into subsystems, each with its own controller. These may then communicate to achieve global objectives. This has been in large part made possible by the advances in ICT and power electronics. Furthermore, the integration of cyber-resources in other systems has been a trend also in other critical systems. The increasing number of *cyber-physical systems*, however, implies the possibility of malicious agents interfering with the operations of critical infrastructure, thus causing significant disruptions, as attested by some high profile cases in recent years.

Moving from this motivation, it is the primary objective of this thesis to define distributed cyber-attack detection architectures capable of detecting the presence of malicious agents: drawing upon a distributed control architecture for DC microgrids as a motivating example, we suppose that attacks may penetrate the communication network included between subsystem controllers to achieve current sharing. Moreover, we endow these distributed diagnosers with plug-and-play capabilities, i.e. the ability of automatically being reconfigured if the system changes its configuration.

These techniques are validated through simulation on DC microgrids, an energy distributed network where generation and consumption of electricity are colocated within a single node, and experimentally on a network of electric springs, a demand-side management technology capable of regulating the voltage and power quality over distribution networks.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Prof Thomas Parisini and Dr Francesca Boem, who have guided me throughout these four years, for their invaluable insight, patience, and their suggestions.

Furthermore, I would like to thank all those professors whom I have had the chance to visit over the years for their generous hospitality, and with whom I have been able to collaborate. Specifically, I would like to thank Prof Giancarlo Ferrari-Trecate at the École polytechnique fédérale de Lausanne for hosting me on two separate occasions, Prof Josep Guerrero at Aalborg University, and Prof S. Y. Ron Hui, whom I have had the pleasure of visiting at the University of Hong Kong. I would also like to thank Prof Davide M Raimondo, for his help during the first steps of my research life.

I would like to thank my examiners, Prof Bikash Pal and Prof Angelo Alessandri, for their valuable comments on this work, which have improved its quality.

I would have not been able to get through this had it not been for a number (a large number) of people who have always supported me, through thick and thin, pushing me in times of difficulties. Firstly, to my friends back home, to Sergio, Dami, Appi, Ghera, Pelle, Franca, Sferu, Banda and Baba. Thank you for making Italy a place I've always wanted to come back to, for the crazy times and excellent discussions, and all the beer and tea we have shared together. To my (extended) Italian Community away from home, who welcomed me at the start of my PhD, Alberto, Antonio, Filiberto, Paolo, Giordano, Thulasi: I have thoroughly enjoyed our time together, and am tremendously grateful to know you. To Nick, Tunvez, Jeevan, Sihao, Sophie, Margs and the rest of the Dindins crew: between the food, the drink, the climbing, and frankly just the hanging out – life's been good. To Mustafa, for accepting my countless ramblings and for welcoming me in Lausanne, together with Harsh and Pulkit. To my friends in CAP, who were always around for a good laugh. To the Yummy bunch, Domenico, Federica, Benita, Marco, Tommaso, Matteo, Adair, for lightening up the mood daily, even since we've been apart. To Mirko, thanks for always being around for a chat, and helping me work through stuff. To Angelo, I thank you both for your friendship, listening to my bs, and for our discussions on work. To the members of e.quinox, past and present, for just being awesome.

More thanks are due to the people that have been around since my arrival at Imperial in undergrad. To Mark, for being the second goonie in our somewhat-functional household, for

the late-night conversations, and countless other shenanigans. To Omar, for being the best adventure-bud I could ever ask for, pushing me in situations that are uncomfortable and yet so incredibly satisfying, as well as his support in the day-to-day. To Jov and Beppe, for their immense friendship.

A special thanks goes to my family, who have supported me through all of my decisions, and who have been there for me throughout. I am particularly thankful to my sister, Sofia, for having shared some of my time in London: it was great to have a bit of home with me.

A final thanks goes to Chiara. Your patience, our laughs, and the time we spend together is more important than you know. I cherish you always.

To Chiara

To my grandparents, for their example

‘We made the world we’re living in
and we have to make it over’

James Baldwin

‘Dov’è Jones il suonatore
Che fu sorpreso dai suoi novant’anni
E con la vita avrebbe ancora giocato’

Fabrizio De André

Contents

Abstract	5
Acknowledgements	7
1 Introduction	21
1.1 Motivation	21
1.2 Problem formulation	23
1.3 Literature review	28
1.4 Contributions	39
1.5 Thesis outline	41
2 Modeling DC microgrids and cyber-attacks – Some background	43
2.1 DC microgrids	43
2.2 Modelling cyber-attacks in a control-theoretic framework	55
2.3 Conclusions	69
3 Distributed cyber-attack detection	71
3.1 Problem formulation	73
3.2 Structural detectability for distributed cyber-attack diagnosers	79
3.3 Local diagnoser $\mathcal{D}_i$ – A multi-module design	87
3.4 Local state observer – $\mathcal{O}_i^L$	93
3.5 Neighbour state observer – $\mathcal{O}_{ji}^{UIO}$	101

3.6	Detectability properties of $\mathcal{D}_i$	114
3.7	Isolability of attacks in the communication network	119
3.8	Plug-and-play design of $\mathcal{D}_i$	123
3.9	Case study – DC microgrids	125
3.10	Conclusions	135
4	Distributed cyber-attack detection under stochastic uncertainties	139
4.1	Problem formulation	142
4.2	Decentralized state and coupling estimation	146
4.3	Detection strategy	156
4.4	Detection filter for neighbouring subsystems	162
4.5	Simulation results – DC microgrids	169
4.6	Conclusions	173
5	Experimental results: networks of electric springs	175
5.1	Networks of Electric Springs	176
5.2	Distributed cyber-attack detection	184
5.3	Experimental results	191
5.4	Conclusions	194
6	Conclusion	197
6.1	Summary of Thesis Achievements	197
6.2	Future Work	199
	Bibliography	200
A	State estimation in the presence of unknown inputs	221
A.1	Unknown inputs: Observability and reconstruction	222
A.2	Unknown-input observers	224
A.3	Minimum-variance unbiased estimator with unknown inputs	234
A.4	Conclusion	240
B	Discretisation of distributed generation unit dynamics	241
C	Augmenting DGU state space for UIO detectability	245

List of Tables

2.1	Bias injection attacks – Attack resources	62
2.2	Zero-dynamics attacks – Attack resources	65
2.3	Replay attacks – Attack resources	67
2.4	Covert attacks – Attack resources	68
3.2	Information required for design of $\mathcal{D}_i$ and attack detection	92
3.3	Interpretation of replay and covert attacks as LTI system (3.93)	117
3.4	Electrical parameters of DC microgrid	129
3.5	Parameters of distributed controllers	129
4.2	Electrical parameters of the DC microgrid	169
5.1	Parameters of network of electric springs	195

List of Figures

1.1	DC microgrid as network of DGUs	24
1.2	Large-scale system as network of subsystems	25
1.3	Stages of plug-and-play operations for LSS	26
2.1	Schematic representation of DGU i	45
2.2	Circuit diagram of Buck converter	50
2.3	Attack resource space	57
2.4	Cyber-physical system subject to an attack	58
3.1	Large-scale system as interconnected subsystems	74
3.2	Distributed diagnoser $\mathcal{D}_i$	88
3.3	Stages of plug-and-play operations for LSS	124
3.4	DC microgrid as network of DGUs	126
3.5	Scenario 1.a.: Coordinated attacks for stealthiness – module $\mathcal{O}_4^L$	132
3.6	Scenario 1. a.: Coordinated attacks for stealthiness – modules $\mathcal{O}_{24}^{UIO}$ and $\mathcal{O}_{34}^{UIO}$.	133
3.7	Scenario 1.b. Structural stealthiness – Residuals and thresholds of $\mathcal{O}_4^L$	136
3.8	Scenario 1.b. Structural stealthiness – modules $\mathcal{O}_{24}^{UIO}$ and $\mathcal{O}_{34}^{UIO}$	137
3.9	Scenario 2 Covert attack against $\mathcal{O}_{24}^{UIO}$	138
4.1	LSS as a network of subsystems	143
4.2	Distributed diagnoser $\mathcal{D}_i$ – Stochastic framework	147
4.3	Distributed diagnoser $\mathcal{D}_i$ – Particular of $\mathcal{O}_i^d$	154

4.4	DC microgrid as network of DGUs	169
4.5	Covert attack against $\mathcal{O}_4^d$	172
4.6	Covert attack against $\mathcal{O}_{24}^d$ and $\mathcal{O}_{34}^d$	172
4.7	Effect of the covert cyber-attacks α_{24} and α_{34} on the DC microgrid as a whole. Plots of each state for all DGUs $i \in \mathcal{N} = \{1, \dots, 6\}$	173
5.1	Network of electric springs	177
5.2	Circuit diagram of electric springs	177
5.3	Park decomposition of electric springs voltage	179
5.4	Diagram of the distributed controller which is regulating ES_i	184
5.5	Electric spring and its diagnoser	189
5.6	Experimental setup	191
5.7	A picture of the experimental setup	192
5.8	Network of electric springs	193
B.1	Circuit diagram of Buck converter	242

Notation

$\mathbb{N}$	Set of natural numbers
$\mathbb{Z}$	Set of integer numbers
$\mathbb{N}_0$	Set of non-negative integers
$\mathbb{R}$	Set of real numbers
$\mathbb{R}^n$	n -dimensional real space
$\mathbb{R}_{>0}^n$	n -dimensional positive orthant
$\mathbb{C}$	Set of complex numbers
$\mathbb{C}^+$	Open right half complex plane
$\mathbb{C}^-$	Open left half complex plane
$x \in \mathcal{A}$	An element x of set $\mathcal{A}$
$\mathcal{A} \subseteq \mathcal{B}$	Set $\mathcal{A}$ is a subset of set $\mathcal{B}$
I_n	The n -dimensional identity matrix; when clear from context, n is omitted
0_n	The n -dimensional matrix of zeros; when clear from context, n is omitted
$x \in \mathbb{R}^n$	An n -th dimensional real vector x
$A \in \mathbb{R}^{n \times n}$	An n by n dimensional real matrix A
$\sigma(A)$	The spectrum of matrix A
Hurwitz-stable	A matrix such that $\sigma(A) \subseteq \mathbb{C}^-$
Schur-stable	A matrix such that $\max \sigma(A) \leq 1$
$A \succ 0$	A symmetric matrix $A = A^\top$ such that $\sigma(A) \subset \mathbb{C}^+$
$A \succcurlyeq 0$	A symmetric matrix $A = A^\top$ such that $\sigma(A) \subseteq \mathbb{C}^+$
$x(T_-)$	Limit from below of a signal $x(\cdot)$
$\mathcal{V}^\perp$	Given a subspace $\mathcal{V}$, $\mathcal{V}^\perp$ is its orthogonal complement

$ \cdot $	Operator $ \cdot $, when applied to a matrix A or vector x , is the component-by-component absolute value; when applied to a set $\mathcal{X}$, it defines its cardinality
$\lceil \cdot \rceil$	The ceiling function
$\lfloor \cdot \rfloor$	The floor function
$\leq, \geq$	Inequalities; when applied to vectors or matrices, the inequality is considered to be taken component-by-component
$\text{Im}A$	The image of matrix $A \in \mathbb{R}^{n \times m}$, i.e. $\{x \in \mathbb{R}^n x = Ab, \forall b \in \mathbb{R}^m\}$
$\ker A$	The nullspace of matrix $A \in \mathbb{R}^{n \times m}$, i.e. $\{x \in \mathbb{R}^n Ax = 0\}$
$\text{rank } A$	The rank of matrix A
$\text{col}_{i \in \mathcal{N}}[x_i]$	The column concatenation of vectors x_i , with indexes i in set $\mathcal{N}$
$\text{row}_{i \in \mathcal{N}}[A_i]$	The row concatenation of matrices A_i , for all $i \in \mathcal{N}$
$\text{diag}_{i \in \mathcal{N}}[a_i]$	The diagonal concatenation of scalars a_i , for all $i \in \mathcal{N}$
$\text{diag}_{i \in \mathcal{N}}[A_i]$	The diagonal concatenation of matrices A_i , for all $i \in \mathcal{N}$
$M = m_{ij}$	A matrix M with m_{ij} as its (i, j) -th component
$M = [M_{ij}]$	A block-matrix M with M_{ij} as its (i, j) -th block
$\doteq$	Is defined as
$\equiv$	Is equivalent to
LC filter	Electronic circuit composed of an inductor and a capacitor
RLC filter	Electronic circuit composed of an inductor, a resistor and a capacitor
DC	Direct current
AC	Alternating current

Chapter 1

Introduction

1.1 Motivation

Over the course of the last few decades, electrical energy grids have seen a surge in distributed generation capacity, with a substantial proportion of that being composed of renewable energy sources [JRRP10, CKFL15] in the challenge to tackle human-made climate change. Renewable energy sources are inherently intermittent and cause bi-directional active power flow across energy distribution lines. Thus, novel methods for voltage and power regulation have been introduced in the literature, including distributed control architectures to guarantee power quality and reliable operations [APKLGH17]. This has been made possible given the advancements in information and communication technologies (ICT) and in power electronics, creating the possibility of effectively and efficiently controlling distributed generation [YBG11].

This evolution of the control paradigm for energy distribution networks has been in lock-step with a more general move in control systems broadly towards the widespread integration of “cyber” resources within physical systems. This has been possible due to an overall reduction in costs associated with microprocessors, sensors and communication devices, thus allowing for a higher penetration of computational, sensing and communication capabilities.

Although digital control systems are nothing new, the sheer scale of this integration has led researchers to dub these systems as “cyber-physical systems” (CPS) to highlight the importance of both their physical and “cyber” properties [BG11, GSC⁺17]. The concept of CPS has spurred

an interest in a number of different problems, such as networked control [Bul18], distributed optimization [NNC⁺19], and a general interest in the so-called “Internet of Things” [Sta14].

Even though it provides significant benefits, the introduction of cyber-resources into control systems has also lead to the possibility of malicious agents which may wish to destabilise, degrade or otherwise harm control systems. Indeed, there are a number of recent high-profile cases where cyber-physical systems were targeted by cyber-attacks [SM07, Cas16, FMC11, HF⁺18, Sob19], thus evidencing the urgency of tackling cyber-security from a system-theoretic perspective [CAS08b]. Specifically, the two incidents described in [Cas16, Sob19] targeted the Ukrainian and U.S. energy networks, causing black-outs and impacting society greatly, further highlighting the need to equip these systems with automatic methods for their security [WS19].

Until recently, however, cyber-security has been evaluated from the point of view of information technology (IT), focussing on data *confidentiality*, *integrity* and *availability*¹, rather than on control theoretic methods for security in cyber-physical systems. It has been pointed out, however, that infrastructure security may be improved by designing methods to mitigate the impact of attacks *after* they have been successfully leveraged, rather than preventing them [CAS08a], by exploiting known physical characteristics of CPS. This has led to the field of *secure control*, which has recently attracted significant levels of interest in the literature (see for example the recent special issues [CSS17, CPH⁺20], as well as the survey [UGC⁺16], the tutorial [CST19], and the monograph [WOM⁺19]).

Another important modelling framework that can be used to describe energy distribution networks and other safety critical infrastructure, apart from that of CPS, is that of large-scale systems (LSS) [Sil90, Lun92]. Having been studied for several decades, LSS are characterised by their scale and complexity, and therefore centralised architectures may not be suitable for their control. Indeed, given their large number of states and their spatial distribution, information aggregation in a single control centre may not be feasible, both for its cost and the computational resources it may require. This has led to the development of non-centralised control architectures.

The focus of this thesis, therefore, is the development of tools and methods to tackle dis-

¹This triad, abbreviated in CIA, is in fact fundamental when dealing with data breaches.

tributed architectures for the detection of cyber-attacks for large-scale cyber-physical systems. Specifically, although throughout the work we often abstract away from their specific properties, the main motivating application for this work is that of smart energy distribution networks. Indeed, the problem formulation presented in Section 1.2 is directly derived from analysing the secondary consensus-based control protocol for current sharing in DC microgrids, presented in [TMGFT18].

Through the integration of power electronics devices to perform demand matching, smart distribution networks are fundamental for the future power grid to be able to manage the ever increasing renewable energy generation capacity being introduced [CKFL15, JRRP10]. These systems are designed to be inherently distributed, require communication for coordination and optimal performance (see for example [CBCZ16] for a discussion of the need of communication to solve the problem of voltage regulation), and are prime targets for cyber-attacks. As a matter of fact, any disturbance to the power grid may cause severe disruptions to society at large, as electrical grids are a critical backbone of modern social systems and are expected to be reliable. Furthermore, the integration of communication networks for regulation of distributed generation units in the future power grid, possibly enabled via wireless devices with limited defence mechanisms [YBG11], offers malicious agents an ideal *entry point* to perform cyber-attacks if countermeasures are not taken to address this [WS19].

1.2 Problem formulation

Let us now formulate the problem addressed throughout this thesis; namely, distributed cyber-attack detection for energy distribution networks, and predominantly for DC microgrids. These are analysed through the modelling framework of large-scale interconnected systems, to provide generality to the results. In Figure 1.1 we offer a visualisation of a DC microgrid as a large-scale interconnected system composed of a number of distributed generation units (DGUs) as subsystems.

In this section, we want to introduce the reader to the scenario considered, and to show how we propose to tackle the problem, rather than to mathematically present the problem formulation (which is defined at the beginning of each chapter).

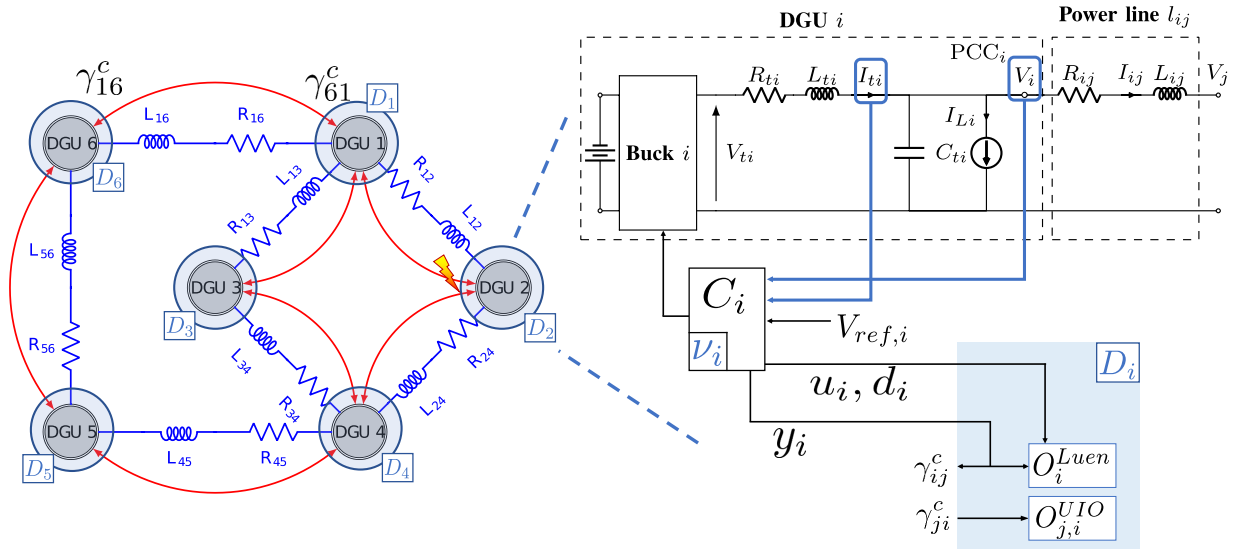


Figure 1.1: DC microgrid represented as a LSS with $N = 6$ DGUS as subsystems. [On the left] DC microgrid – each node represents a DGU, in blue the physical power lines between subsystems, in red the bi-directional communication links. [On the right] Representation of a distributed generation unit (composed of a power source, a Buck DC-DC converter, and an RLC filter), its controller unit C_i and the diagnostic unit D_i .

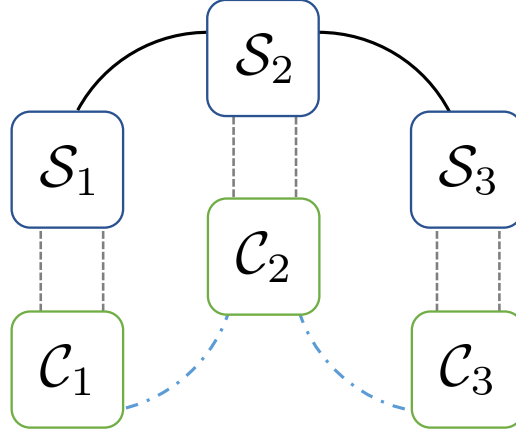


Figure 1.2: Representation of a large-scale system as a network of $N = 3$ subsystems, equipped with controllers $\mathcal{C}_i, i \in \{1, 2, 3\}$. The solid black lines indicate physical coupling between subsystems; the dotted grey lines are the measurement and actuation channels between subsystems $\mathcal{S}_i$ and their controllers $\mathcal{C}_i$; the dashed-dotted blue lines represent the communication between subsystems.

Inspired by DC microgrids, we suppose that large-scale systems are divided according to a non-overlapping decomposition into a set of physically interconnected subsystems [Lun92]. Each of these subsystems is regulated by its own local controller. We consider a cooperative distributed control law, and therefore suppose that each subsystem exchanges information with a subset of all other agents to achieve some common objective. We call this set of subsystems the set of *neighbours* of a subsystem $\mathcal{S}_i$. In Figure 1.2 we give a basic pictorial representation of this architecture.

In view of our previous discussion, we consider that malicious agents are able to penetrate the communication network between subsystem controllers, thus altering the information transmitted. Therefore, a main objective is to design a distributed detection architecture capable of detecting these attacks, by including distributed diagnosis units $\mathcal{D}_i$. Each unit is tasked with monitoring the information received from all neighbours, and detecting whether they have been maliciously altered. We exploit a *model-based* approach inspired by those in the fault detection and isolation (FDI) literature: namely, by exploiting the knowledge of the subsystem's dynamical model, we equip $\mathcal{D}_i$ with residual generators to detect attacks.

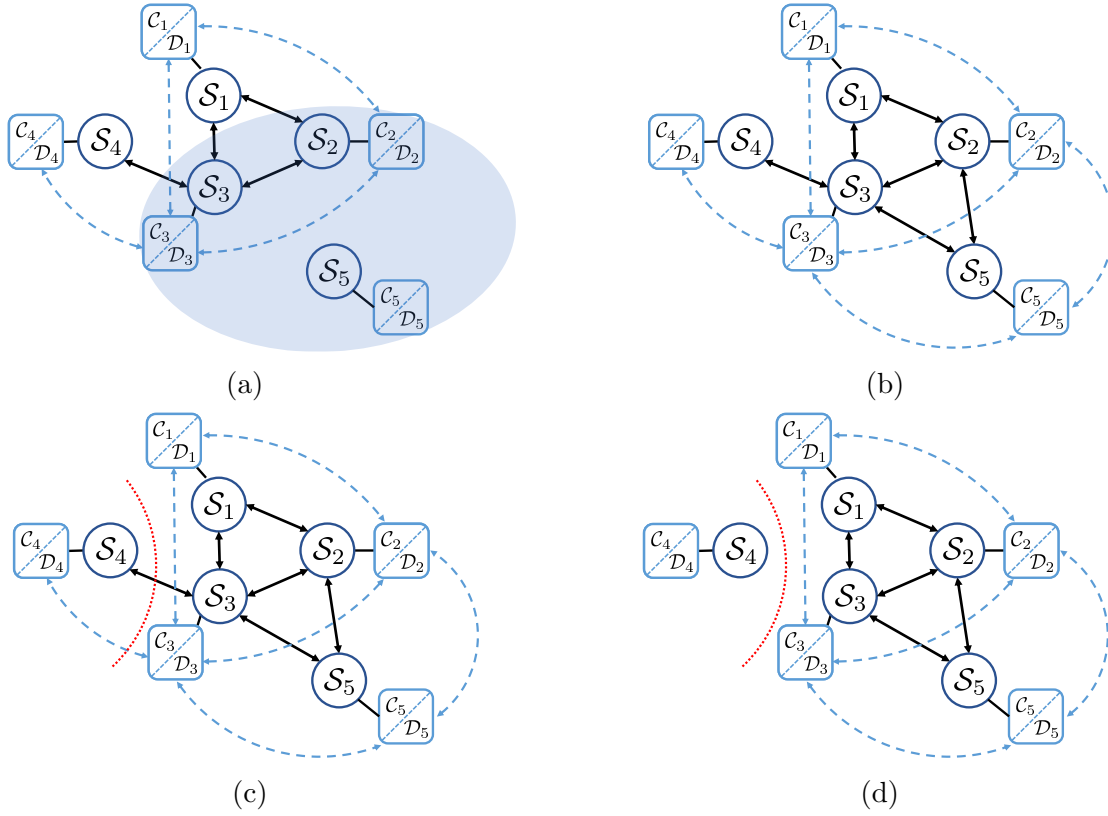


Figure 1.3: Different stages of plug-and-play operation of a large-scale system: (a) S_5 before plug-in, $N = 5$ – neighbours of S_5 , $\mathcal{N}_5 \doteq \{2, 3\}$, and subsystem to be plugged-in are highlighted; (b) LSS after plug-in of S_5 , $N = 5$; (c) LSS before unplugging S_4 ; (d) S_4 unplugged from LSS.

Furthermore, we provide architectures with plug-and-play characteristics. Formally defined in the following, plug-and-play control [Sto09, BTS13, RFFT13, BL15] refers to the capability of distributed controllers to automatically be reconfigured after a subsystem is added or removed from the LSS as a whole, i.e. “plugged-in” and “unplugged”, while not needing all regulators to be reconstructed. This is particularly important in future power networks, as any change to the overall system, from topological changes to subsystems that may be added or removed, must not jeopardise the overall stability of the system, and should not require all controllers to be reconfigured. This framework has also been extended to distributed monitoring architectures, such as the fault detection and isolation methods proposed in [BRFTP15, BRFTP19]. We provide an example of plug-and-play stages in Figure 1.3.

The scenario we consider here, abstracted to the language of LSS, is primarily based on hierarchical control methods for DC microgrids [MST⁺17]. Specifically, as proposed in [TRFT17, TMGFT18], the microgrid is considered as a network of distributed generation units

(DGUs), regulated to achieve current sharing and voltage balancing through a control architecture composed of a decentralised voltage-stabilising primary controller, and a distributed consensus-based secondary controller to achieve current sharing among the DGUs. This distributed control architecture requires the integration of a communication network between subsystems, in order to allow for coordination between them, therefore exposing these systems to cyber-attacks. This control architecture therefore motivates the development of the detection methods proposed in this thesis. Note, furthermore, that the controllers proposed in [TRFT17, TMGFT18] are endowed with plug-and-play capabilities, which is necessary for our methods to be plug-and-play as well.

Let us explicitly note that throughout this thesis we do not consider it possible for an attacker to penetrate the signals that are exchanged between any subsystem $\mathcal{S}_i$ and its controller $\mathcal{C}_i$, assuming that they are colocated. This is, once more, motivated by the case of DC microgrids. Indeed, in each DGU, the controller is colocated with the sensors and actuators of the DGU itself, as they usually lie on the same chip; thus, any malicious agent would need to have compromised the controller to be able to alter this information. As is seen in the remainder of this thesis, this constitutes a critical assumption for the architectures that we propose.

Throughout this work, by “distributed architecture” we intend any architecture for which the design and operation rely on information available locally, and, at most, on that from the neighbouring subsystems. Let us clarify this with an example: suppose the LSS is composed of three subsystems, such as in Figure 1.2. For this system, the set of neighbours² of subsystem $\mathcal{S}_2$ is $\{1, 3\}$, while subsystems $\mathcal{S}_1$ and $\mathcal{S}_3$ only have $\mathcal{S}_2$ as their neighbour. We say that the controllers $\mathcal{C}_i$ for $i \in \{1, 2, 3\}$ are distributed only if:

1. they have been designed with *a priori* model knowledge of the subsystem $\mathcal{S}_i$ they are regulating, and at most that of their neighbours;
2. during online operations, they only require information from their sets of neighbours to achieve the desired control objective.

This is equivalent to $\mathcal{C}_1$ only having access to information about $\mathcal{S}_1$ and $\mathcal{S}_2$ during both design and operation, i.e. not knowing anything about $\mathcal{S}_3$.

²In what follows, sets of subsystems are indicated by the sets of their indexes, to simplify notation.

We note that the definition of the term “distributed” we consider here differs from some of the literature on distributed estimation, and on control-theoretic cyber-security [WM17, TSJ10], where information of the overall system model is available *a priori*, while limiting information in real time to the subsystem itself and its neighbours.

When analysing the effectiveness of our architectures, we consider that malicious agents can violate the integrity of one or more communication links, thus being able to inject malicious signals into the communication. Furthermore, we compare our performance to attackers with varying degrees of resources. As we summarise in Chapter 2, it was shown in [TSSJ15] that the more resources an attacker may leverage (the amount of model knowledge they have, and what level of disruption they can cause), the harder it may be to detect.

1.3 Literature review

Let us now present a general overview of the literature. Starting with an outline of non-centralised control architectures for large-scale and cyber-physical systems, we present the work on plug-and-play control, delineating its main characteristics and properties. Following this, we focus on fault-tolerant and secure control: fields of inquiry where research is done to increase the safety and reliability of anomalous behaviour, such as that caused by faults and cyber-attacks. In our summary of this literature, we focus on distributed architecture for fault and cyber-attack *detection*. Furthermore, we provide an overview of methods in the secure control literature addressing cyber-security of smart grids and, specifically, DC microgrids.

1.3.1 Control of Large-Scale and Cyber-Physical Systems

As mentioned previously, large-scale systems (LSS) describe those systems in which a large number of states interacts over an area which may be large [Lun92, Sil90, SP11]. Cyber-physical systems (CPS), on the other hand, are defined in [BG11] as a “new generation of systems with integrated computational and physical capabilities that can interact with humans”, in turn implying that the components within the system as a whole interact with each other either

physically, through communication and other *cyber* links, or both. Both LSS and CPS can be subdivided into multiple subsystems interacting with each other [SP11].

Over the last decade, a significant number of different fields have addressed the problem of CPS [GSC⁺17], attesting to the cross-discipline nature of these systems. Indeed, in [GSC⁺17], a “survey-of-surveys” on security in CPS, a number of different applications are identified, including smart grids, intelligent transport systems, industrial control systems, medical devices, and manufacturing, among others. Challenges regarding cyber-physical systems were identified by [Zan17], in the introduction to a special issue of *Computer* treating CPS, highlighting how survivability and reliability, and the constraints of power consumption of the embedded systems, are fundamental characteristics for CPS. Furthermore, as pointed out by [Ser18], additional challenges are identified in the development of various levels of abstraction to manage the integration of cyber-resources within physical systems, which scale from embedded systems to industrial process controllers, as well as posing numerous business, regulatory and ethical challenges [Ser18]. In the editorial to a recent special issue [DAAW20], *smart* CPS are addressed, collecting works in different fields, from data processing to security to cyber-enabled smart applications, defining a “next generation of engineering systems” capable of leveraging collaboration to remove boundaries between the physical and virtual world. Although the vision described in [DAAW20] may still not be realisable, a number of different research communities are striving to achieve ever more seamless interaction.

The control system community “has been consistently engaged in the development of theories, tools, and practices for the design and operation of [cyber-physical] systems” [JPTT14]. Indeed, the special issue [JPTT14] covered problems of abstraction and verification, security, resource-constrained embedded and wireless control, and event-based estimation and control. Over the past years, these fields have received increased attention, as attested by [LKZ⁺16, CSS17] and the works cited therein. Further detail on secure control for cyber-physical systems is given in Section 1.3.4.

The scale and interconnections which lie at the core of the definition of large-scale and cyber-physical systems, as well as the possibility of decomposing them into less complex subsystems, has lead to the development of non-centralised control architectures, as LSSs must be divided

to be controlled, as pointed out by [Bak08].

In [SVAS78], the nature of centralised control architectures is identified in the requirement that all information necessary for regulation must be collected in a centralised location. There, a controller can compute the regulating input, to be transmitted to all the available actuation resources. Hence, centralised architectures require point-to-point communication structures [Yan06]. In turn, this implies that the centralised controller must be capable of dealing with the complexity of the LSS in real time, and that the communication network be cost-effective and/or feasible.

Decentralised control architectures were theorised in response to these technical hurdles in the '70s, as shown by the survey [SVAS78], and have since gathered a large amount of interest in the control theoretic community, as demonstrated by [Sil90, Lun92, Bak08, Bak14, DHWG19], and the works cited therein, with large success in many fields: especially notable is the case of smart grids, where distributed control techniques have gained significant traction [APKLGH17, MST⁺17].

Decentralised control design relies on dividing the overall system into *non-overlapping* partitions, i.e. into subsystems in which no state is shared³ [Lun92], and then regulating each subsystem independently of all others. Thus, the physical coupling with neighbouring subsystems is considered as a disturbance. While this allows for control of the overall system without the need of communication between subsystems, it comes at the cost of performance and at times the impossibility of stabilizing the entire system.

The concept of distributed control addresses this issue, by considering the interconnection of a subsystem with its neighbours, which may be known *a priori*, and introducing a communication network linking separate controllers to share information. This however leads to added computational complexity and the need for communication. Clearly, both the complexity and cost of the communication vary from a fully connected distributed architecture to a partially connected one. As LSS and CPS are often weakly coupled [Lun92, Sca09, Bak14, DHWG19], partially connected architectures are often considered sufficient.

³If subsystems are allowed to contain the same states, the partition is defined *overlapping* – for a work in which both overlapping and non-overlapping partitions are exploited, see [RBFTP14].

It is worth noting that the inclusion of a communication network in the control architecture introduces delays to the system, due to the properties of the network itself [Yan06]. While this may not be a problem in general, depending on the application there are scenarios in which this may alter the stability properties of the system as a whole, such as inverter synchronization in low-inertia power networks [MDH⁺18]. However, the importance of communication between subsystems to achieve some “higher level” objectives, such as power sharing in these power networks, often leads to design methods satisfying multiple objectives in a *hierarchical* fashion. These are often referred to as either “hierarchical” or “multi-level” control strategies [FBB⁺80, Lun92, MMT00, MST⁺17]. In hierarchical architectures, different *tiers* of controllers are designed: primary control often offers decentralised stabilization of the network, while higher level controllers guarantee some global performance objectives throughout the network as a whole.

Hierarchical control schemes have also been implemented in the context of “multi-agent systems”, where they are exploited to implement cooperation between subsystems (agents) [Mur07]. One of the problems that has been focused on in the cooperative control literature is the concept of *consensus*, for which multiple agents achieve common agreement on the value of a state [OSFM07, Sha08]. This may be useful in diverse contexts, such as coordination of multi-vehicle formations and flocking [OS06, RBA07], sensor fusion in wireless sensor networks [OS05], and power sharing [ZD15, MDRP⁺16, TMGFT18, CYY⁺19]. Foundational theory behind consensus-based control has been derived from graph theory [OSM04], where the use of Laplacians allows for the analysis of both stability and the speed of convergence of the consensus state.

Finally, it is worth noting that the terms “decentralised” and “distributed” have also been used to define the controller design procedure itself [BL88]. Specifically, decentralised design has been used to define when each controller is computed “completely independently” from its neighbours’ regulators. On the other hand, distributed design refers to a synthesis strategy which takes advantage of the information available from the subsystem’s neighbours [RFSFT13]. These design strategies prove useful when the LSS has a level of complexity which cannot be handled in a single off-line regulator design, as well as in scenarios where multiple parties hold

intellectual property rights on the subsystems composing the LSS, or other issues of privacy may be relevant, and, therefore, the system as a whole may be unknown. This is the case in Internet-of-Things applications, as well as in microgrids [GCLL13].

1.3.2 Plug-and-play Control

Within the literature on non-centralised control architectures, plug-and-play control addresses the design and operational issues which may occur when subsystems are added or removed from the LSS or CPS [BTS13, Sto09, RFSFT13, BL15]. The plugging-in or unplugging of subsystems may be seen as the replacement of a component in a manufacturing process, or the addition of a renewable energy unit in a distribution network, or a *smart device* connecting or disconnecting from the internet, following activation by a user [Sto09, GBMP13].

In this framework, regulator design is performed such that stability properties of the overall system are not spoiled either by adding or removing a subsystem from the overall system. Furthermore, these actions do not require all the controllers in the LSS to be recomputed, but rather only those of the neighbours of the subsystem [RFSFT13, RBFTP16]. Specifically, [RFSFT13] defines *plug-and-play capabilities* as:

- information required in the design of a local plug-and-play controller comes from the subsystem being regulated, and at most from its neighbours;
- an automatic procedure is activated verifying overall system stability in the event of a subsystem being plugged-in/unplugged, which may include controller redesign for the neighbouring subsystems.

It is worth noting, additionally, that a growing number of works have focused on decentralised and distributed plug-and-play control strategies for the regulation of DC microgrids [TRFT17, SSK17, TMGFT18, NST⁺20], addressing both the problem of voltage stabilization and balancing across a network of DGUs, as well as current sharing. Further details on the plug-and-play control architectures presented in [TRFT17, TMGFT18] are given in Chapter 2.

Plug-and-play has also been found as effective within Fault-Tolerant Control, where it may be used to reconfigure the controllers once a fault is detected [KPU06, RBFTP16, BL15,

RBGP16]. In the case of [RBFTP16, RBGP16], once a fault is detected, the faulty subsystem is unplugged from the rest of the LSS, and its neighbours' controllers are reconfigured to ensure stability. On the other hand, in [BL15] plug-and-play reconfiguration is used to redesign the controller of a faulty subsystem.

1.3.3 Fault-Tolerant Control

As large-scale and cyber-physical systems are used to describe processes and systems the failure of which may cause great damage to people, things, or severe economic losses, an objective in the control community has been to develop techniques capable of diagnosing and mitigating the presence of faults, leading to the field of Fault-Tolerant Control (FTC), as “faults in a single component have major effects on the availability and performance of the system as a whole” [BKL⁺06]. This field has been developed over the course of several decades, as attested to by the survey [Pat97], the recent book [BKL⁺06] (which has reached its 3rd Edition in 2016), and the references therein.

It is useful to recall the definition given in [IB97] of the terms “fault” and “failure”, which delineate specific system conditions in the FTC literature, although these may be interchanged in daily discourse:

- A *fault* is “an unpermitted deviation of at least one characteristic property or parameter of the system from the acceptable / usable / standard condition”;
- A *failure* is “a permanent interruption of a system’s ability to perform a required function under specified operating conditions”.

The overall goal of FTC can be seen, then, as an attempt to prevent a fault from becoming a failure, i.e. to act upon the malfunctioning system before a possibly permanent performance degradation occurs [BKL⁺06]. Specifically, as defined in [BKL⁺06], FTC “concerns the interaction between a given system and a controller” which does not only include the nominal control law, but also “analyses the behaviour of the plant in order to identify faults and changes the control law to hold the closed-loop system in a region of acceptable performance”. Again,

relying on [BKL⁺06] for definitions, “acceptable performance” can be seen as related to the notions of *safety*, *reliability*, *availability* and *dependability*.

Many fault-tolerant control techniques have been developed over the years, which has therefore inspired different categorizations of their methods. One such distinction, seen in [JY12], is between *passive* and *active* approaches. *Passive* FTC refers to controller design which is robust to possible faults, and therefore is capable of maintaining performance even in the presence of a fault. This is achieved at the cost of optimality of the control in nominal conditions. On the other hand, *active* FTC techniques develop automatic protocols that the controller follows once it is “aware” of a fault. This is the class of fault-tolerant control design techniques seen in the remainder of this section.

In *active* fault-tolerant controller design, the term “controller” is used to describe not only the regulation loop used to govern the system, but a series of decision making tools designed to detect whether the system has a faulty component, and to determine the reconfiguration of the system once a fault has been detected [BKL⁺06]. The first of these two problems is defined *fault diagnosis*, while the latter is known as *controller redesign*.

To achieve fault diagnosis, the system must be equipped with a monitoring layer capable of determining whether the behaviour of the system is nominal, i.e. not affected by a fault, or if it is faulty, and then to identify the entity and scale of the fault. Fault diagnosis is therefore composed of two separate steps: one to detect the presence of a fault, and the second to isolate the fault affecting the dynamics of the system.

Once the fault diagnosis procedure has successfully identified the fault acting upon the system, the control law can be reconfigured to account for it, thus returning the system to acceptable performance levels, whilst minimizing control switching transients. By exploiting this methodology within the class of active techniques, suboptimal operation in nominal conditions can be avoided.

A large body of work has been dedicated to fault detection and isolation (FDI) starting from the 1970s [Wil76], as can be seen in the surveys [Fra96, GF97, IB97, GCD15], books [Ise06, Din08, BKL⁺06], and the works cited therein. A classification of different methodologies for FDI, which can be quantitative, qualitative, or process history based, can be found in

[VRK03, VRKY03, VRYK03]. The first are based on quantitative models of the systems to be monitored, and expected behavior is compared to measured behavior, so that it is possible to evaluate the inconsistency between the model and the real system. Qualitative methods, on the other hand, rely on a description of the process based on causal relationships, an abstraction of the system's interactions, and the plant's performance is analyzed within this framework in order to detect faults. Finally, process history methods rely on the availability of large amounts of process data, rather than information on the underlying structure of the system, which is then evaluated to detect malfunctions.

Although this large body of work exists in the context of FDI, much of it focuses on centralised monitoring architectures, and therefore many of the results may be unsuitable for FTC techniques tailored to large-scale or cyber-physical systems. However, decentralised and distributed architectures for FDI methods have been developed, of which [PKC⁺07, SID⁺10, STSJ11, BFP11, FPP12, AMP15, BKLS16, BSS16, DMK16, RBFTP16, RBGP16, BFK⁺17, BRFTP19, KZC⁺20] are all examples. These works form the foundation of the distributed techniques employed for cyber-attack detection in this thesis.

1.3.4 Secure Control

Similarly to fault-tolerant control, the idea behind secure control is to monitor systems which may be exposed to attacks, i.e. the influence of malicious agents attempting to alter the nominal behaviour of the system. This has become a relevant feature of control architectures, attracting lots of attention in the literature, as can be seen from the recent surveys and special issues [UGC⁺16, SAJ15, CSS17, GUC⁺18, CST19, CPH⁺20] and references therein. As identified in [WS19], the problem of secure control for cyber-physical systems may be split, with detection, identification, and resilience all having their role to play. In the remainder of this section, and throughout this thesis, we focus primarily on the issue of detection of cyber-attacks in CPS.

Cyber security, as well as information security, is a field that has received a great deal of attention in the computer science and information technology (IT) literature [vSvN13]. However, secure control of CPS and LSS presents several differences, due to the specificities of CPS and IT systems. Thus, different approaches must be defined when dealing with secure

control [CAS08b], since the software for the regulation of CPS cannot be frequently patched, CPSs and LSSs must be available in real time to avoid grave repercussions, and are often made up of legacy systems. Furthermore, in Secure Control, the physical interconnection between components can be exploited to evaluate whether attacks are active on the system's network.

Similarly to cyber security, the objectives of Secure Control are to maintain the *confidentiality*, *integrity*, and *availability* of the LSS or CPS, which are also called the CIA triad [vSvN13, CAS08a]. Confidentiality refers to the possibility of maintaining the data/information inaccessible to external, unauthorised entities; integrity highlights the trustworthiness of the data which is transmitted over the network, i.e. that it not be tampered with; finally, availability indicates the ability of the system to be online when required.

To achieve these objectives, methods of attack detection and resilience towards attacks have been focused on [CAS08b]. These are related to the concepts of fault diagnosis and control reconfiguration discussed earlier for FTC. However, there are some crucial differences between faults and cyber-attacks, namely how they are constructed.

In fact, although both faults and attacks can be modelled as an external signals influencing the dynamics of the system, the magnitude and activation time of which are unknown, faults are generally considered to be random in nature, whilst attacks are the result of calculation and may be coordinated over time and space as a decision of the attacker [PBB12, TSSJ15]. This in turn leads to difficulties in detection: indeed, based on disruption and disclosure resources available to the attacker, together with the knowledge it has of the CPS model, cyber-attacks may be designed explicitly to be *stealthy* (i.e. undetectable) to the monitoring scheme [TSSJ15].

Different types of attacks have been classified. For example, [CAS08a] defines the following attack strategies:

- *disclosure attacks*, in which the malicious agent monitors information transmitted over a system's communication network without authorization;
- *deception attacks*, where false information is injected in the control loop, possibly affecting the output measurements or the inputs transmitted over the network;

- *Denial of Service* (DoS) or *jamming* attacks, where the transmission is blocked, thus compromising the availability of the system;
- direct attacks on the physical plant, not in the “cyber” space.

A further taxonomy is provided by [TSSJ15], where the authors emphasise the resources that are required by an attacker to successfully achieve a stealthy attack, defining the “cyber-physical attack space” over the following “coordinates”: *disclosure resources*, i.e. the possibility of carrying out a disclosure attack; *disruption resources*, related to how the attack influences the dynamics of the system; *model knowledge*. Classic examples of cyber-attacks targeting control systems are the STUXNET virus [MWS15], the Maroochy water breach [SM07], as well as the attack on the Ukrainian power grid [Cas16]. More detailed information on the modelling of these cyber-attacks in cyber-physical systems is presented in Chapter 2, where we focus on disclosure and deception attacks.

It was recently highlighted in [POM⁺18] that attack detection and resilience algorithms can be categorised by dividing them into *data-driven* and *knowledge-based* approaches. We here focus on the latter, as they share similarities to model-based FDI and FTC, and form the basis of the algorithms designed in this thesis. A consistent body of literature on knowledge based techniques is now available, mostly focusing on centralised architectures to detect malicious intrusion and tampering with the communication network connecting controllers and plants, e.g. [TSSJ15, PDB13, Smi15], among others.

While a number of works evaluates *passive* detection [PDB13, TPSJ12, BPG17, FT20], it is worth noting that *active* attack detection methods have also been proposed, such as [MWS15, WS15], where the control input to the CPS are altered to improve the possibility of detection, at the cost of suboptimally controlling the system.

Lately, the interest in algorithms treating the distributed case has grown, as can be seen from the publications [PDB13, NI14, DPA⁺18, PDB15b, FBC⁺17, DUSL19, AKP20, BRBP20]. Furthermore, the recent special issue [CPH⁺20] focused specifically on distributed algorithms for security and privacy of networked systems.

Of the works cited above, [NI14] and [DPA⁺18] consider the case in which diagnosers have

access to complete system knowledge, but only a subset of the system’s outputs. [PDB15b] proposes a distributed algorithm for attack identification, while supposing communication between controllers to be secure, i.e. not subject to attacks. In [FBC⁺17], a distributed detection and secure set estimation algorithm is developed, based on random set theory; here, the problem is formulated considering the communication between subsystem nodes to be secure. A distributed detection and mitigation strategy in distributed estimation is proposed in [DUSL19], where it is assumed that the global model of the system is available to each estimating node. [BRBP20], focusing on distributed detection of attacks that are locally covert, again suppose the cyber-attack diagnosis modules of each subsystem can communicate securely. Finally [AKP20] compare centralised and decentralised detection strategies in the context of interconnected systems subject to stochastic uncertainties, analysing the strengths and weaknesses of both.

Specific attention has also been given to smart grids [ME10, MKB⁺12, PDB13, SHG⁺12, WL13, MMM20], and particularly to microgrids [ZYBV15, DSM⁺16, FIDSDB17, BNJD18, SMPD18, ZMQ⁺18, LLZC19], motivated by a recognition from governments and practitioners that security in power networks is fundamental [NER09], and the fact that “cyber security [alone] is not sufficient to secure the smart grid” [WS19]. Indeed, the smart grid is a system which is too complex and heterogeneous to remove all possible “attack surfaces”. Therefore, [WS19] argues for the need of CPS security, comprised of three main objectives: attack detection, isolation of the components under attack, and the design of resilient algorithms to maintain acceptable levels of performance even in the presence of attacks.

Of the works addressing security for microgrids, [ZMQ⁺18] focuses on assessing the impact of false data injection attacks within microgrids, analysing their effect on a load sharing protocol. Instead, the authors of [BNJD18, SMPD18, LLZC19] address the problem of detecting cyber-attack in microgrids. Specifically, *Signal Temporal Logic* (STL) is exploited in [BNJD18] to detect whether an attack is present, verifying whether some given STL conditions are violated. On the other hand, “balanced” attacks are considered in [SMPD18], defining a Cooperative Vulnerability Factor (CVF), and design their scheme such that each diagnoser can exploit secure information of each of its neighbouring subsystems. Finally, [LLZC19] considers a primal-dual distributed optimization algorithm for power sharing among the generation units, and exploit

the dual variable for detection.

1.4 Contributions

As mentioned in Section 1.2, the main objective of this thesis is to design a distributed diagnosis architecture for cyber-attack detection in large-scale cyber-physical systems, with DC microgrids as a motivating case study. Indeed, similarly to the case of DC microgrids, we consider a set of interconnected subsystems regulated through a distributed control architecture to achieve some common objective (such as current sharing), and consider an attacker capable of disrupting the information exchanged across the communication network required to enable cooperation.

The main contribution of this work is the presentation of a distributed cyber-attack detection architecture for large-scale cyber-physical systems. Although generalised to a wide class of systems, the proposed methodologies are strongly related to power-electronics-enabled future energy distribution networks, and specifically the application of DC microgrids. In fact, the problem formulation and some of the results presented were found by analysing this specific test case. We analyse the strengths and, more importantly, the weaknesses of this architecture in the presence of unknown-but-bounded disturbances and stochastic uncertainties.

Moreover, we present a fundamental limitation to the detectability of cyber-attacks for distributed detection architectures. Indeed, we prove that there are certain scenarios, depending on the physical coupling between subsystems, where limiting the information available for design of the diagnoser $\mathcal{D}_i$ to the sole dynamics of the local subsystem $\mathcal{S}_i$ may strongly limit the cyber-attack-detection properties of $\mathcal{D}_i$. This is indeed the case for DC microgrids, where the physical coupling is not sufficiently *rich* to enable a local attack diagnoser $\mathcal{D}_i$ to detect any attack on the variables used for the consensus-based control scheme to achieve current sharing.

Finally, we demonstrate that the proposed detection architecture is effective, through both simulations using realistic models of DC microgrids, as well as experimentally, on networks of electric springs. The latter were carried out in the Power Electronics Laboratory in the Electrical and Electronic Engineering Department of the University of Hong Kong, through the generous hospitality of Prof Shu Yuen Ron Hui, whom we thank.

This thesis draws upon the following published work:

- [GTB⁺20] A. J. Gallo, M. S. Turan, F. Boem, T. Parisini, G. Ferrari-Trecate, “A Distributed Cyber-attack Detection Scheme with application to DC Microgrids”, *IEEE Transactions on Automatic Control*, vol. 65, n. 9, pp. 3800–3815, 2020.
- [BGFTP17] F. Boem, A. J. Gallo, G. Ferrari-Trecate, T. Parisini, “A distributed attack detection method for multi-agent systems governed by consensus-based control”, *56th Annual Conference on Decision and Control (CDC)*, pp. 5961–5966, 2017
- [GTN⁺18] A. J. Gallo, M. S. Turan, P. Nahata, F. Boem, T. Parisini, G. Ferrari-Trecate, “Distributed cyber-attack detection in the secondary control of DC microgrids”, *European Control Conference (ECC)*, pp. 344–349, 2018.
- [BGBP19] A. Barboni, A. J. Gallo, F. Boem, T. Parisini, “A distributed approach for the detection of covert attacks in interconnected systems with stochastic uncertainties”, *58th Annual Conference on Decision and Control (CDC)*, pp. 5623–5628, 2019
- [GBP20] A. J. Gallo, A. Barboni, T. Parisini, “On detectability of cyber-attacks for large-scale interconnected systems”, *21st IFAC World Congress*, 2020, *To appear*

Furthermore, during the course of the PhD, the following have been published, although the results are not directly related to this thesis:

- [BGRP20] F. Boem, A. J. Gallo, D. M. Raimondo, T. Parisini, “Distributed fault-tolerant control of large-scale systems: an active fault diagnosis approach”, *IEEE Transactions on Control of Network Systems*, vol. 7, no. 1, pp. 288–301, 2020

- [GTB⁺] A. J. Gallo, M. S. Turan, F. Boem, G. Ferrari-Trecate, T. Parisini, “Distributed watermarking for secure control of microgrids under replay attacks”, IFAC-PapersOnLine, vol. 51, n. 23, pp.182-187, 2018

1.5 Thesis outline

The rest of the thesis is structured as follows. In Chapter 2 we summarise background material used throughout the thesis: specifically, we present the modelling framework of DC microgrids, including their dynamical modelling, and the plug-and-play hierarchical control architecture considered. Furthermore, we give a brief introduction to results on cyber-attacks in cyber-physical systems, relying heavily on results proposed by [TSSJ15]; looking at the problem from a centralised point of view, the concept of “attack space” is introduced; we then introduce some well-known classes of attacks.

In Chapter 3 we present the proposed distributed cyber-attack detection architecture for a large-scale system with continuous-time dynamics, affected by unknown-but-bounded uncertainties. After formulating the problem mathematically, we offer structural limitations of distributed monitoring schemes when restricting the information available to the diagnoser $\mathcal{D}_i$; motivated by this, we present a multi-module architecture, which generates residuals for attack detection exploiting model knowledge of the local subsystem and its neighbours. We provide in-depth analysis of the properties of the diagnoser constructed in this way. Moreover, we demonstrate how to isolate the communication link over which an attack is active, and show how the proposed design satisfies the requirements to have plug-and-play characteristics. Finally, we show the effectiveness of the detection strategy on a realistic model of a DC microgrid.

In Chapter 4, we propose a diagnosis tool to detect cyber-attacks for systems subject to stochastic uncertainties, discretising the dynamics of the large-scale system. We analyse the statistical properties of this diagnoser, illustrating how it can be designed to be fully distributed, even without any approximations. Even in this case, we show the effectiveness of the diagnosis module by presenting simulations on a DC microgrid.

Having presented these distributed frameworks, and having shown their effectiveness in simulation, in Chapter 5 we turn our attention to experimental results. Indeed, after presenting the experimental setup considered, i.e. a network of electric springs, we adapt the scheme proposed in Chapter 4, to give evidence of the real-world applicability of the proposed cyber-attack detection architecture.

In Chapter 6 we present concluding remarks, indicating some possible future directions for this research. We propose an overview of some methods which may be explored to accommodate the effect of cyber-attacks on distributed control protocols automatically.

In the appendices, we give more in-depth background information on the estimation schemes adapted in this thesis [Appendix A], and give some supplementary considerations on DC microgrids: we provide a discussion on the discretisation of DC microgrids accounting for the switching nature of their switch-mode power supply [Appendix B], and an augmented state is proposed to allow for detection by unknown-input observers [Appendix C].

Chapter 2

Background

Having presented the motivation for secure control in cyber-physical systems, and having formulated the problem to account for the specificities of DC microgrids, let us now present some background necessary for the proper comprehension of the remainder of this thesis. Specifically, in this chapter we give an overview both on how to model DC microgrids, and how cyber-attacks are modelled in a control theoretic framework.

2.1 DC microgrids

Microgrids, both AC and DC, are spatially distributed power networks in which energy generation, storage and consumption are often colocated. Indeed, microgrids offer advantages over conventional power networks when dealing with renewable energy sources, as they can increase the power quality and reduce the network losses. For these reasons, interest in microgrids in the literature has greatly increased over the past decade [GVM⁺10, ZD15, MST⁺17]. DC microgrids prove especially appealing, as renewable energy sources, energy storage technology, and a growing number of consumer loads are inherently DC in nature. Specifically, our focus here is on low-voltage DC microgrids, which find application in data-centers, smart houses, and electric vehicle charging, among others [MST⁺17].

DC microgrids can be seen as being composed of a number of distributed generation units, interconnected through electrical distribution lines, and composed of a DC voltage source, a

microgrid interfacing converter and a load. Controller architectures presented for the regulation of DC microgrids are often hierarchical in nature: a decentralised controller within each DGU regulates the voltage, guaranteeing voltage stability for the DC microgrid as a whole [TRFT17, NST⁺20, ZD15]; distributed controllers provide secondary regulation to achieve global system objectives, such as improving power quality, enabling current/power sharing among DGUs, often exploiting consensus-like protocols to adjust the set-point of the primary controller [MDRP⁺16, TMGFT18, ZD15]; further tertiary control is often designed to perform system-wide optimization and management centrally. A detailed survey of hierarchical control methods for DC microgrids can be found in [MST⁺17].

Given the nature of the hierarchical controllers designed for the operations of DC microgrids, the control system must be endowed with a communication network to achieve them [CBCZ16, MST⁺17]. Coupled with the fact that power networks are a primary target for attack (see for example the recent attack on the Ukrainian power network [Cas16, LAC16] and the more recent Denial of Service attack against the U.S. grid [Sob19]), the integration of communication networks within the control system exposes DC microgrids to cyber-attacks. As such, they offer an ideal case study for the distributed cyber-security framework presented in this thesis.

In the remainder of this section, we present a framework to model DC microgrids, together with the primary and secondary controllers which are considered in Chapter 3.9. Furthermore, we discuss some of the difficulties related to implementation of the system in simulation, such as the switching nature of the input to the system.

2.1.1 Dynamical model of a DC microgrid

In this thesis, similarly to [TRFT17], we consider the DC microgrid as a large-scale interconnected system, composed of N physically coupled distributed generation units (DGUs). DGUs are composed of a voltage source V_s (such as a PV panel or a battery), a Buck DC-DC converter with terminal voltage and current V_{ti} and I_{ti} , an RLC circuit connecting each DGU to the point of common coupling (PCC) with voltage V_i , and a current sink I_{Li} , representing the load. The circuit diagram of a DGU is given in Figure 2.1. Each DGU i is interconnected at the point of

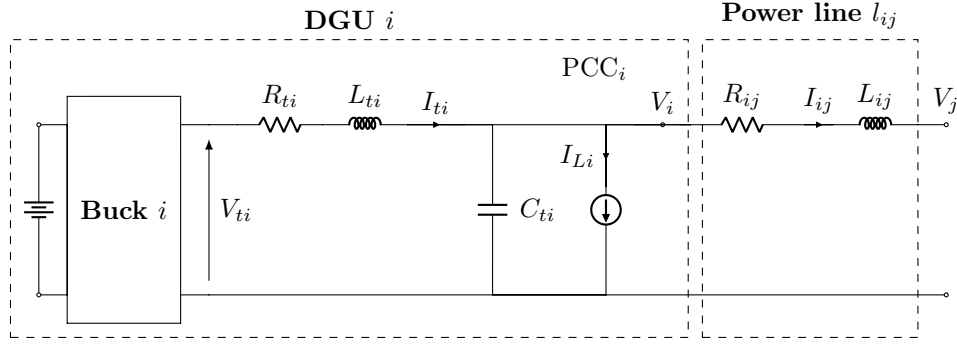


Figure 2.1: Schematic representation of DGU i , power line l_{ij} and interconnection with PCC voltage of DGU j . Note, power line l_{ij} is represented with both its resistive and inductive components.

common coupling to a set $\mathcal{N}_i$ of neighbouring DGUs through power lines l_{ij} . We assume that the following conditions hold:

Assumption 2.1. *The following hold for all DGU i in the DCmG:*

1. *The source voltage $V_i^s > V_i$;*
2. *V_i^s is constant ;*
3. *The local load is connected at the point of common coupling ;*
4. *The Buck converter always operates in continuous conduction mode, i.e. $I_{ti}(t) > 0, \forall t \geq 0$.*

◁

The dynamics of each DGU are defined by the linear time-invariant dynamic equations:

$$\begin{cases} C_{ti}\dot{V}_i = I_{ti} - I_{Li} + \sum_{j \in \mathcal{N}_i} I_{ij} \\ L_{ti}\dot{I}_{ti} = -V_i - R_{ti}I_{ti} + V_{ti} \end{cases} \quad (2.1)$$

where V_i is the PCC voltage, I_{ti} is the terminal current of the Buck converter; together, V_i and I_{ti} are the state of DGU i . The terminal voltage V_{ti} is the control input to the subsystem, while load current I_{Li} is an exogenous input. The electric parameters of the RLC circuit are R_{ti} , L_{ti} , and C_{ti} . Currents $I_{ij} = -I_{ji}$ are the current through the power line l_{ij} , which link DGU i to a set of its neighbours¹, which we define $\mathcal{N}_i \doteq \{j \in \mathcal{N} | R_{ij} < \infty\} \subseteq \mathcal{N} \doteq \{1, \dots, N\}$, where R_{ij}

¹Note that, in the context of DC microgrids, $j \in \mathcal{N}_i$ if and only if $i \in \mathcal{N}_j$, i.e. the graph induced by the physical interconnections between DGUs is undirected.

is the resistance of line l_{ij} . The dynamics of the line current I_{ij} are:

$$L_{ij}\dot{I}_{ij} = -V_i - R_{ij}I_{ij} + V_j \quad (2.2)$$

with L_{ij} the inductance of line l_{ij} . Reproducing the modelling framework of [TRFT17], we consider the following assumption.

Assumption 2.2 (Quasi Stationary Lines). *For all $i, j \in \mathcal{N}$, line inductances are negligible, i.e. $L_{ij} \approx 0$.* $\triangleleft$

Given Assumption 2.2, the dynamic equation (2.2) is substituted with the following algebraic expression:

$$I_{ij} = \frac{V_j - V_i}{R_{ij}} \quad (2.3)$$

and as such the voltage dynamics in (2.1) are equivalent to:

$$C_{ti}\dot{V}_i = I_{ti} - I_{Li} + \sum_{j \in \mathcal{N}_i} \frac{1}{R_{ij}} (V_j - V_i). \quad (2.4)$$

2.1.2 Plug-and-play distributed control for DCmG

We now summarise the control strategy presented in [TRFT17, TMGFT18]. The control objectives are threefold: guarantee overall voltage stability of the DC microgrid; maintain voltage balancing; and achieve current sharing. Let us formally define these objectives.

Definition 2.1 (Voltage balancing). *The PCC voltages of the DGUs are said to achieve voltage balancing if:*

$$\frac{1}{N} \sum_{i \in \mathcal{N}} V_i = V_{ref} \quad (2.5)$$

for some common reference value V_{ref} . $\triangleleft$

Definition 2.2 (Current sharing). *For constant load currents, DGUs are said to achieve current sharing if:*

$$\lim_{t \rightarrow \infty} \left| \frac{I_{ti}}{I_{ti}^s} - \frac{I_{tj}}{I_{tj}^s} \right| = 0 \quad \forall i, j \in \mathcal{N} \quad (2.6)$$

for given scaling factors $I_{ti}^s > 0, \forall i \in \mathcal{N}$. $\triangleleft$

Remark 2.1. Note that the scaling factors I_{ti}^s are introduced to allow for the possibility of different ratings of each DGU. If this is the case, in fact, we do not require them to contribute the same value of current, but rather the same proportion, possibly measured in per unit (p.u.). Furthermore, if all scaling factors are the same throughout all DGUs, condition (2.6) is equivalent to

$$I_{ti} = \frac{1}{N} \sum_{i \in \mathcal{N}} I_{Li}, \quad (2.7)$$

for all $i \in \mathcal{N}$, i.e. the average of the load currents shared amongst all DGUs [TMGFT18]. $\triangleleft$

To achieve these objectives, a hierarchical controller is designed, composed of primary and secondary control loops. The primary control loop guarantees voltage stability, while the secondary loop achieves current sharing and voltage balancing. Furthermore, the primary controller is designed to be decentralised, i.e. to only rely on local information, while the secondary controller is designed to be distributed, i.e. it relies on information at most from DGUs neighbouring DGU i . The design of both controllers is such that they are plug-and-play², i.e. after DGU i is added to (plugged in) or removed from (unplugged) the DC microgrid, only the controllers of those DGUs neighbouring DGU i must be redesigned. In the following we introduce these controllers, but refer the interested reader to [TRFT17, TMGFT18] for extensive analysis of their properties.

Primary controller

The primary controller guarantees voltage stability of the DGU when interconnected with other DGUs in a DC microgrid, and introduces voltage tracking capabilities.

To ensure the local voltage V_i tracks the reference $V_{ref,i}$, an integral state ν_i is added, with dynamics:

$$\dot{\nu}_i = V_{ref,i} - V_i. \quad (2.8)$$

The control input V_{ti} is then computed as the following state feedback law:

$$V_{ti} = k_{i,1}V_i + k_{i,2}I_{ti} + k_{i,3}\nu_i \quad (2.9)$$

²Recall discussion in Chapter 1.3.2.

with $K_i \doteq [k_{i,1}, k_{i,2}, k_{i,3}]$ designed such that the closed loop DC microgrid is stable [TRFT17]. In [TRFT17], K_i is found by solving an appropriately defined linear matrix inequality (LMI). Additionally, further analysis in [NST⁺20] has shown that a decentralised stabilising K_i can be designed for each DGU also when Assumption 2.2 does not hold, achieving voltage reference tracking. Moreover, the authors of [NST⁺20] define the following inequalities on the components of K_i , which are sufficient for (2.9) to be a decentralised feedback controller stabilizing the DC microgrid as a whole:

$$\begin{cases} k_{i,1} < 1; \\ k_{i,2} < R_{ti}; \\ 0 < k_{i,3} < \frac{1}{L_{ti}} (k_{i,1} - 1)(k_{i,2} - R_{ti}). \end{cases} \quad (2.10)$$

Throughout the remainder, we assume the following.

Assumption 2.3. *All DGUs have the same global voltage reference $V_{ref} \geq 0$.* $\triangleleft$

Remark 2.2. It is clear to see that, under Assumption 2.3, the primary controller alone achieves voltage balancing, as all voltages V_i track V_{ref} . $\triangleleft$

Let us now present the secondary controller, which achieves current sharing by altering the local voltage reference.

Secondary controller

The secondary controller is designed to achieve current sharing while preserving voltage balancing, given Assumption 2.3. A distributed consensus protocol is introduced in (2.12), itself requiring a communication network to be included in the control architecture.

The secondary controller does not affect the input of DGU i directly, rather it alters the primary controller's reference. Thus, a local reference voltage is introduced, defined as:

$$V_{ref,i} = V_{ref} + \Delta V_i, \quad (2.11)$$

where a compensation variable ΔV_i is included.

To allow for the possibility of a communication network with topology different from that of the physical DC microgrid, a new set of neighbours is defined, $\mathcal{N}_i^c$, with which DGU i has a communication link. The consensus protocol is then formalised as:

$$\Delta \dot{V}_i = k_I \sum_{j \in \mathcal{N}_i^c} a_{ij} \left(\frac{I_{tj}^c}{I_{tj}^s} - \frac{I_{ti}^m}{I_{ti}^s} \right) \quad (2.12)$$

where the superscripts c and m of variables I_{ti} stand for “communicated” and “measured”, respectively; a_{ij} are the consensus weights, satisfying $a_{ij} = a_{ji}$; finally, $k_I > 0$ is a scaling factor common to all DGUs.

Plug-and-play control design

Summarising the control strategy described above, we rewrite the controller as a linear time-invariant dynamical system, with V_{ti} as its output. Specifically:

$$\begin{cases} \Delta \dot{V}_i = k_I \sum_{j \in \mathcal{N}_i^c} a_{ij} \left(\frac{I_{tj}^c}{I_{tj}^s} - \frac{I_{ti}^m}{I_{ti}^s} \right) \\ \dot{\nu}_i = V_{ref} + \Delta V_i - V_i \\ V_{ti} = k_{i,1} V_i + k_{i,2} I_{ti} + k_{i,3} \nu_i. \end{cases} \quad (2.13)$$

By following the design procedures in [TRFT17, NST⁺20, TMGFT18] the controller (2.13) has plug-and-play capabilities, thus allowing for DGUs to be added and removed from the DC microgrid without spoiling the stability of the overall system.

2.1.3 Switching inputs

Thus far, we have treated the DGU as a linear time-invariant system. In reality, Buck DC-DC converters are Switched Mode Power Supplies (SMPSs), implemented through power electronic switches, such as insulated-gate bipolar transistors (IGBTs), as shown in Figure 2.2. Indeed, the dynamics of the DGU change depending on whether switch D in Figure 2.2 is *open* or *closed*. In the remainder of this section, we redefine the dynamics in (2.1) to include the switching of the SMPS. The following benefited from the discussion presented in [Mar17].

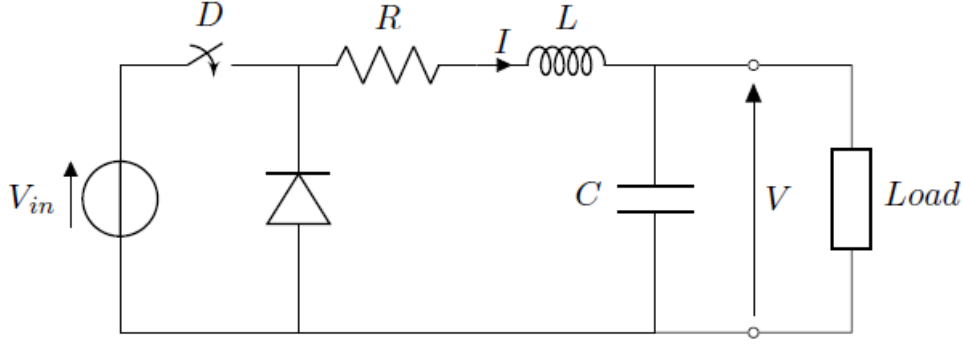


Figure 2.2: Circuit diagram of Buck converter in series with an RLC filter and a resistive load. The switch D represents a transistor, such as an IGBT.

Switching dynamics of the Buck converter

We now present the dynamics of the Buck converters in its two modes of operation during each switching cycle: with switch D open and closed. To simplify notation throughout this section, we drop subscripts i relating to any particular DGU, without risk of ambiguity.

Let us write the dynamics of the converter as the hybrid system:

$$\dot{x} = A_p x + B_p \bar{u} \quad (2.14)$$

where $x \doteq [V, I_t]^\top$ is the state of the system, $\bar{u} \doteq V_s$ is the voltage of the DC power supply, considered constant, and where we have disregarded the load current I_L . Furthermore, $p \in \{O, I\}$ indicates the discrete states in which switch D may be in, with O and I representing the *open* and *closed* states, respectively. Finally, A_p and B_p are appropriately defined matrices, of the form:

$$A_O \doteq \begin{bmatrix} 0 & \frac{1}{C_t} \\ -\frac{1}{L_t} & -\frac{R_t}{L_t} \end{bmatrix} \quad B_O = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (2.15)$$

$$A_I \doteq \begin{bmatrix} 0 & \frac{1}{C_t} \\ -\frac{1}{L_t} & -\frac{R_t}{L_t} \end{bmatrix} \quad B_I = \begin{bmatrix} 0 \\ \frac{1}{L_t} \end{bmatrix} \quad (2.16)$$

assuming the semiconductor components (namely the switch and the diode) are ideal.

During operation, the Buck converter is operated at a nominal switching frequency, $f^{\text{sw}} =$

$1/T^{\text{sw}}$. Over the switching period, the switch is kept closed for some period $T^{\text{on}} < T^{\text{sw}}$, and then opened for the remainder of the switching period $T^{\text{sw}} = 1/f^{\text{sw}}$. We therefore define the duty cycle:

$$\delta \doteq \frac{T^{\text{on}}}{T^{\text{sw}}} \quad (2.17)$$

and its complement

$$\delta^* \doteq 1 - \delta = \frac{T^{\text{off}}}{T^{\text{sw}}}, \quad (2.18)$$

where $T^{\text{off}} \doteq T^{\text{sw}} - T^{\text{on}}$.

In the literature, it is often the case to analyse the “average” dynamics of SMPS, i.e. the dynamics (2.14) averaged over a switching period (see for example [MĆ77]). Specifically, to approximate the dynamics over a single switching period T^{sw} , an “average” state x^{avg} is introduced, with dynamics:

$$\dot{x}^{\text{avg}} = \delta^*(A_O x^{\text{avg}} + B_O \bar{u}) + \delta(A_I x^{\text{avg}} + B_I \bar{u}). \quad (2.19)$$

These represent the “average” of the “off” and “on” dynamics, weighted by the complement of the duty cycle δ^* and the duty cycle δ , respectively. Given that $A_I = A_O = A$ and $B_O = 0$, as can be seen from their definitions in (2.15)-(2.16), the dynamics of x^{avg} simplify to:

$$\dot{x}^{\text{avg}} = A x^{\text{avg}} + \delta B_I \bar{u}, \quad (2.20)$$

where $\bar{u} = V_s$ is considered to be constant.

By rearranging (2.20), a time-varying control input $u(t) = V_t(t)$ can be defined as

$$V_t(t) = \delta(t) V_s \quad (2.21)$$

which, for $\delta(t)$ computed at each switching instant $t = \kappa T^{\text{sw}}, \kappa \in \mathbb{N}_0$, is constant for all $t \in [\kappa T^{\text{sw}}, (\kappa + 1) T^{\text{sw}}), \kappa \in \mathbb{N}_0$.

Considering primary control (2.9) designed in Section 2.1.2, the duty cycle $\delta(t)$ is computed

as:

$$\delta(t) = \frac{1}{V_s} [k_1 V(\kappa T^{\text{sw}}) + k_2 I_t(\kappa T^{\text{sw}}) + k_3 \nu(\kappa T^{\text{sw}})] \quad t \in [\kappa T^{\text{sw}}, (\kappa + 1)T^{\text{sw}}). \quad (2.22)$$

In the following, let us introduce the notation

$$\delta_\kappa = \delta(\kappa T^{\text{sw}}) \quad \text{and} \quad \delta_\kappa^* = 1 - \delta(\kappa T^{\text{sw}}). \quad (2.23)$$

Having defined the average dynamics (2.20), let us analyse the error that is introduced by this modelling framework, i.e. $e = x - x^{\text{avg}}$. Suppose without loss of generality some initial time $t_0 = 0$, and that the initial state for both the switching and average dynamics is defined as $x(0) = x^{\text{avg}}(0) = x_0$. Note that the solution of x^{avg} after one switching period T^{sw} , given dynamics in (2.20), can be found to be

$$\begin{aligned} x^{\text{avg}}(T^{\text{sw}}) &= e^{AT^{\text{sw}}} x_0 + \int_0^{T^{\text{sw}}} e^{A(T^{\text{sw}}-\tau)} \delta_0 B_I \bar{u} d\tau \\ &= e^{AT^{\text{sw}}} x_0 - A^{-1}(I - e^{AT^{\text{sw}}}) \delta_0 B_I V_s. \end{aligned} \quad (2.24)$$

On the other hand, the exact solution of the state at time T^{sw} is:

$$x(T^{\text{sw}}) = e^{A_O \delta_0^* T^{\text{sw}}} e^{A_I \delta_0 T^{\text{sw}}} x_0 + \int_0^{\delta_0 T^{\text{sw}}} e^{A_I(\delta_0 T^{\text{sw}}-\tau)} B_I \bar{u} d\tau + \int_{\delta_0 T^{\text{sw}}}^{T^{\text{sw}}} e^{A_O(T^{\text{sw}}-\tau)} B_O \bar{u} d\tau. \quad (2.25)$$

Recalling that $A_I = A_O = A$ and $B_O = 0$, and $\bar{u} = V_s, \forall t \geq 0$, (2.25) simplifies to:

$$\begin{aligned} x(T^{\text{sw}}) &= e^{AT^{\text{sw}}} x_0 + \int_0^{\delta_0 T^{\text{sw}}} e^{A(\delta_0 T^{\text{sw}}-\tau)} B_I \bar{u} d\tau \\ &= e^{AT^{\text{sw}}} x_0 + A^{-1} (I - e^{A\delta_0 T^{\text{sw}}}) B_I V_s \end{aligned} \quad (2.26)$$

Finally, by taking the first terms of the Taylor expansion of the exponential $e^{Pt} \approx I + Pt$ we see that (2.24) is an approximation of (2.26), as:

$$e^{AT^{\text{sw}}} x_0 - A^{-1}(I - e^{AT^{\text{sw}}}) \delta_0 B_I V_s \approx e^{AT^{\text{sw}}} x_0 + \delta_0 T^{\text{sw}} B_I V_s \approx e^{AT^{\text{sw}}} x_0 + A^{-1} (I - e^{A\delta_0 T^{\text{sw}}}) B_I V_s. \quad (2.27)$$

The average approximation of the dynamics (2.14) as (2.20), therefore, removes the high-order frequency components of the behaviour of the DGU. Thus, it is useful for analysis, so long as f^{sw} is higher than the cutoff frequency of the RLC filter. It must be noted however that this approximation does introduce a non-zero error, which may become an issue when estimating the state of the DGU from real measurements or in realistic simulations. Indeed, having defined the error $e = x - x^{\text{avg}}$, its dynamics are:

$$\dot{e}(t) = Ae(t) + B_I(V_t(t) - \delta(t)V_s), \quad (2.28)$$

which, given stability of A and that $V_t(t) - \delta(t)V_s \neq 0$ in general, implies that

$$\lim_{t \rightarrow \infty} e(t) \neq 0, \quad (2.29)$$

i.e. that the average error converges asymptotically to some non-zero value.

However, by introducing another conceptual approximation, we can model the switching of D as being a property of the input voltage V_t , which therefore becomes a switching signal. Specifically, we define $V_t(t)$, for all $\kappa \in \mathbb{N}_0$, as:

$$V_t(t) = \begin{cases} V_s, & t \in [\kappa T^{\text{sw}}, (\kappa + \delta T(\kappa T^{\text{sw}}))T^{\text{sw}}) \\ 0, & t \in [\kappa T^{\text{sw}} + \delta T(\kappa T^{\text{sw}}), (\kappa + 1)T^{\text{sw}}) \end{cases} \quad (2.30)$$

This approximation is used throughout the remainder of this section, as well as in Chapter 3.9 and Chapter 4.5, where it is exploited to compute an unbiased estimate of the state of the DGU.

2.1.4 State-space description of a DGU

Let us summarise the discussion in the previous sections by describing the state-space model of the dynamics of the DGU, accounting for the switching nature of the terminal voltage of the buck converter, both in continuous and discrete time.

For modelling in continuous time, consider the state of the DGU as $x_i \doteq \begin{bmatrix} V_i & I_{ti} & \nu_i \end{bmatrix}^\top$,

and recall physical dynamics of the DGU in (2.1), the physical coupling between subsystems as in (2.3), and the controller (2.13). Thus, the state-space model of DGU i can be written as:

$$\begin{cases} \dot{x}_i = A_{ii}x_i + B_i u_i^{\text{sw}} + \sum_{j \in \mathcal{N}_i} A_{ij}x_j + E_i d_i \\ y_i = C_i x_i \end{cases} \quad (2.31)$$

with u_i^{sw} the switching control input, $d_i = \begin{bmatrix} I_{Li} & V_{ref,i} \end{bmatrix}^\top$ the exogenous input composed of the local load current and voltage reference, and where the matrices are defined by:

$$\begin{aligned} A_{ii} &\doteq \begin{bmatrix} -\sum_{j \in \mathcal{N}_i} \frac{1}{C_{ti}R_{ij}} & \frac{1}{C_{ti}} & 0 \\ -\frac{1}{L_{ti}} & -\frac{R_{ti}}{L_{ti}} & 0 \\ -1 & 0 & 0 \end{bmatrix}, \quad B_i \doteq \begin{bmatrix} 0 \\ \frac{1}{L_{ti}} \\ 0 \end{bmatrix}, \\ A_{ij} &\doteq \begin{bmatrix} \frac{1}{C_{ti}R_{ti}} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad E_i \doteq \begin{bmatrix} -\frac{1}{C_{ti}} & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}. \end{aligned} \quad (2.32)$$

The switching control input u_i^{sw} is defined as in (2.30), for all $\kappa \in \mathbb{N}_0$:

$$u_i^{\text{sw}}(t) = \begin{cases} V_{si}, & t \in [\kappa T_i^{\text{sw}}, \kappa T_i^{\text{sw}} + \delta_i(\kappa T_i^{\text{sw}})T_i^{\text{sw}}) \\ 0, & t \in [\kappa T_i^{\text{sw}} + \delta_i(\kappa T_i^{\text{sw}})T_i^{\text{sw}}, (\kappa + 1)T_i^{\text{sw}}) \end{cases} \quad (2.33)$$

where V_{si} is the source voltage of the i -th DGU, and the duty cycle δ_i is computed from the primary control loop:

$$\delta_i(t) = \frac{1}{V_{si}} K_i x_i(\kappa T_i^{\text{sw}}), \quad t \in [\kappa T_i^{\text{sw}}, (\kappa + 1)T_i^{\text{sw}}), \quad \kappa \in \mathbb{N}_0 \quad (2.34)$$

with $K_i \doteq [k_{i,1}, k_{i,2}, k_{i,3}]$. The local voltage reference $V_{ref,i}$ is dependent on the secondary control loop, and therefore defined as:

$$V_{ref,i} = V_{ref} + \Delta V_i \quad (2.35)$$

where V_{ref} is the voltage reference common to all DGUs [Assumption 2.3], while ΔV_i is the result of the consensus protocol (2.12). Furthermore, it is assumed that both V_i and I_{ti} are measured, while ν_i is known exactly, as it is internal to the controller. Hence, $C_i \doteq I_3$.

2.2 Modelling cyber-attacks in a control-theoretic framework

Differently from the remainder of this thesis, here we consider the problem from a centralised point of view, to simplify notation, while providing some insight specifically into the modelling framework of cyber-attacks. We also present some specific classes of *stealthy attacks*, i.e. attacks that are designed purposefully to be undetected, to be used in what follows to investigate the performance of our distributed diagnosis architecture.

Borrowing heavily from both the results and the structure of [TSSJ15], we give a description of how these cyber-attacks may be constructed, by analysing their effect on a generic linear CPS. Throughout this chapter, starting from the three-dimensional description of attacker resources proposed in [TSSJ15], we describe the centralised model for the dynamics of a networked cyber-physical system. We then highlight the main characteristics of cyber-attacks in networked control scenarios, focusing on their differences with respect to faults. Finally, we define some specific classes of attacks.

2.2.1 Attacker resources and knowledge

Let us start by highlighting the main characteristics of cyber-attacks in cyber-physical systems, differentiating them from how cyber-attacks are considered in other IT systems. Furthermore we point out the characteristics that distinguish them from faults, which are more familiar in the control-theoretic literature and have been studied extensively.

From a control-theoretic perspective, the main concern is that cyber-attacks leveraged against a cyber-physical system may influence the physical dynamics of the plants by acting upon the communication networks through which the control systems *close the loop*. This

is not dissimilar to faults, which may alter the way the plant or the actuators/sensors behave, thus causing at least one characteristic property of the system to deviate unacceptably from its nominal value: indeed, recall the definition of the term “fault” that was given in Chapter 1.3.3. The anomalies caused by either cyber-attacks or faults, then, may affect the physical security of system operators, the general public, and the infrastructure alike, as well as influence the economic performance of the system.

However, differently from faults, cyber-attacks offer a different type of challenge to ensure the safe and secure operation of control systems. Indeed, faults are usually considered to be malfunctions of different components within the feedback system, such as the actuators, the sensors, or the plant itself. Therefore, they are often modelled as random, non cooperating events, and a common assumption is that there is ever only one fault active at a time on a plant. Conversely, cyber-attacks are distinct both in their capability of multiple attacking agents colluding to achieve a common objective, and in their malicious intent [TSSJ15].

Throughout the remainder of this chapter we present different classes of attacks, taking the taxonomy presented in [TSSJ15], classified according to the following three characteristic properties, as shown in Figure 2.3:

1. *a priori* model knowledge available to the malicious agent, i.e. a basis with which to describe the amount of model knowledge required for an attacker to pursue a given attack strategy;
2. the amount of disclosure resources that can be leveraged to gather data in real-time, i.e. the amount of online information an attacker has access to, such as real time actuation and measurement signals, by violating their confidentiality;
3. the disruption resources used to alter the behaviour of the system, i.e. the potential of an attacker in disrupting the system’s operations.

Note that in [TSSJ15] both data and physical resources are considered to be potentially disrupted. However, we are only interested in the attacker’s ability of altering data transmitted over communication channels, and, as such, no changes to the physical behaviour of the system is considered in what follows.

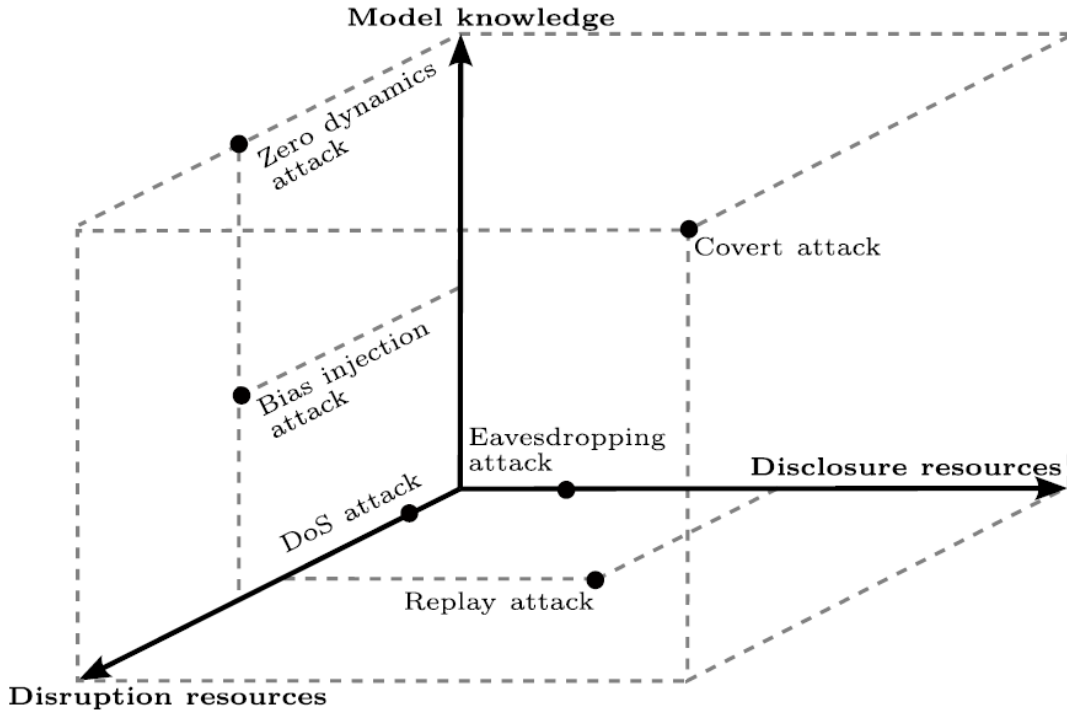


Figure 2.3: Attacker resource space based on the three dimensions proposed in [TSSJ15], from which the figure is taken.

By focusing on the properties of cyber-attacks in this framework, it is possible to ascertain the “difficulty” for a malicious agent to perform different types of attacks. Indeed, the larger the “norm” in the attack space, as can be seen in Figure 2.3, the more capabilities (in terms of either information or resources) the attacker must possess. Although the exact placement of the various classes of attacks in the graph shown in Figure 2.3 is clearly qualitative, it is nonetheless useful to think of an attacker capable of performing a *covert attack* as being more *sophisticated*³ than one capable of implementing a *denial of service* (DoS) attack. Therefore, the above three properties can be seen as a proxy for the degree of sophistication possessed by the attack.

2.2.2 Modelling cyber-attacks in Cyber-Physical Systems

Modelling Cyber-Physical Systems

We consider a general centralised model of the CPS including the physical dynamics of the plant $\mathcal{S}$, its dynamic controller $\mathcal{C}$, and an anomaly detector $\mathcal{D}$, as shown in Figure 2.4. A CPS

³By “sophistication” here we loosely mean the capability of an attacker to engage in complex attacks which may have a large impact on the CPS while at the same time being hard to detect.

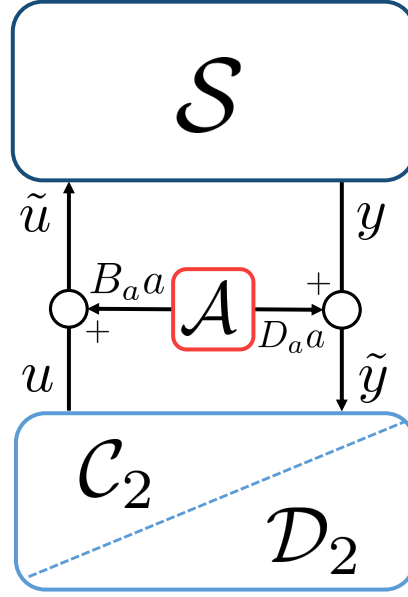


Figure 2.4: Diagram of a cyber-physical system under attack. A plant $\mathcal{S}$ is equipped with a controller $\mathcal{C}$ and a diagnoser $\mathcal{D}$. The input and output signals are transmitted over a communication network, and may be altered by a malicious agent $\mathcal{A}$.

can be modelled as a LTI system with the following state-space representation:

$$\begin{cases} \dot{x} = Ax + B\tilde{u} + Ed + w \\ y = Cx + v \end{cases} \quad (2.36)$$

where $x \in \mathbb{R}^n$, $\tilde{u} \in \mathbb{R}^m$, $y \in \mathbb{R}^p$ are, respectively, the state, the input to the actuators and the sensor measurements; vector d represents external inputs; while w and v model process and measurement disturbances, respectively. Here, we consider them to be bounded noise terms, satisfying

$$|w(t)| \leq \bar{w}(t), \quad |v(t)| \leq \bar{v}(t) \quad \forall t \geq 0 \quad (2.37)$$

with $\bar{w}(t)$, $\bar{v}(t)$ known by the defender for all $t \geq 0$.

We consider a dynamic feedback controller $\mathcal{C}$, which may be designed to achieve different objectives, such as stability or optimal performance, that can be described as a linear time-invariant output feedback controller:

$$\begin{cases} \dot{z} = A_c z + B_c \tilde{y} \\ u = C_c z + D_c \tilde{y} \end{cases} \quad (2.38)$$

with $z \in \mathbb{R}^{n_z}$ the internal state of the controller, and A_c , B_c , C_c and D_c , describing its dynamics, of appropriate dimensions.

Moreover, similarly to those designed for fault detection and isolation (see for example [MVW89]), the diagnoser⁴ $\mathcal{D}$ includes a residual generator, which can be described as the following LTI dynamical system:

$$\begin{cases} \dot{\zeta} = A_d \zeta + B_d u + K_d \tilde{y} \\ r = C_d \zeta + D_s u + F_d \tilde{y} \end{cases} \quad (2.39)$$

where $\zeta \in \mathbb{R}^{n_\zeta}$ is the internal state of $\mathcal{D}$, and $r \in \mathbb{R}^{p_r}$ is the residual. We consider that the residual generator (2.39) is designed to detect cyber-attacks on the CPS, as shown in the following. Furthermore, in the remainder of this chapter, we suppose that the closed loop dynamics composed of (2.36) and (2.38), as well as the dynamics (2.39) are stable by construction of the controller and the residual generator.

Remark 2.3. Note that, in (2.36), (2.38) and (2.39), we use both $\tilde{u}$, $\tilde{y}$ and u , y to represent the input and output signals of the system. This is done to highlight the presence of a communication network between the plant $\mathcal{S}$ and its controller and diagnoser units $\mathcal{C}$ and $\mathcal{D}$ over which the measurement and control-input signals are transmitted, as shown in Figure 2.4. Specifically, $\tilde{u}$ is the control input as transmitted to the system $\mathcal{S}$ by the controller, while $\tilde{y}$ is the measurement output signal as received by $\mathcal{D}$ and $\mathcal{C}$ from the plant $\mathcal{S}$. It is over this communication network that we consider malicious agents to be capable of injecting attacks in this chapter. $\triangleleft$

Modelling cyber-attacks

Attacks on the communication network can be modelled as additive signals on the channels transmitting y and u between the plant and the controller/diagnoser. Specifically, they can be seen as an agent $\mathcal{A}$ computing vector $a(t)$

$$\tilde{u} = u + B_a a(t) \quad (2.40a)$$

⁴Throughout the text “anomaly detector”, “diagnoser”, and “detection module” are used interchangeably.

$$\tilde{y} = y + D_a a(t) \quad (2.40b)$$

where B_a and D_a are such that

$$\text{Im}B_a \subseteq \text{Im}B, \quad \text{Im}D_a \subseteq \text{Im}C, \quad (2.41)$$

i.e. the attack is capable of affecting input signals transmitted to actuators and measurement signals communicated from sensors. The design of the attack signal $a(t)$ is dependent on *a priori* knowledge and operational resources possessed by the attacking agent $\mathcal{A}$, as well as the attack strategy that is pursued.

Definition 2.3 (Nominal behaviour). *The plant of a CPS is said to be operating nominally at time τ if, for all $t \in [0, \tau]$:*

$$a(t) = 0, \quad \tilde{y}(t) = y(t), \quad \tilde{u}(t) = u(t) \quad (2.42)$$

where $y(t)$, $u(t)$ are the unaltered output measurement and control input. $\triangleleft$

Remark 2.4. Here we only consider attacks on the communication network between the plant $\mathcal{S}$ and the controller/diagnoser modules $\mathcal{C}$ and $\mathcal{D}$ and thus the representation of attack vector as an additive signal is consistent with the literature on fault detection and isolation (FDI). Note that, as noted in [MVW89], multiplicative effects on the plant $\mathcal{S}$ can also be seen as additive signals on the plant dynamics of the form of (2.40), possibly with $\text{Im}B_a \not\subseteq \text{Im}B$ or $\text{Im}D_a \not\subseteq \text{Im}C$. $\triangleleft$

Given our interest in designing an attack diagnosis architecture for large-scale CPS, this can be expressed in the objective of detecting and isolating cyber-attacks which may be active on the communication network. Let us therefore formally define “detection” and “isolation”, borrowing terminology from the FDI literature (see, for example, [ZPP02]).

Definition 2.4 (Attack detection). *An attack is said to be detected if a monitoring unit $\mathcal{D}$ establishes that system $\mathcal{S}$ is no longer operating in nominal behaviour.* $\triangleleft$

Definition 2.5 (Attack isolation). *The monitoring unit $\mathcal{D}$ is said to isolate an attack if it is able to establish the type and/or the location of an attack active on $\mathcal{S}$.* $\triangleleft$

Furthermore, we define two properties that are attributed to attacks, which offer insight in the properties of the diagnoser $\mathcal{D}$:

Definition 2.6 (Detectable attacks). *An attack signal $a(t)$ is said to be detectable by a diagnosis module $\mathcal{D}$ if it is such that it is guaranteed to be detected.* $\triangleleft$

Definition 2.7 (Stealthy attacks). *An attack signal $a(t)$ is said to be stealthy to a diagnosis module $\mathcal{D}$ if it is designed such that $\mathcal{D}$ is guaranteed to not detect it.* $\triangleleft$

Although the specific design of $\mathcal{C}$ and $\mathcal{D}$ is not of interest at this stage, we nevertheless provide an overview of how $\mathcal{D}$ may be designed to detect attacks. Suppose that the residual generator (2.39) is defined so that an appropriate threshold $\bar{r}(t)$ may be defined, satisfying

$$|r(t)| \leq \bar{r}(t) \quad (2.43)$$

in nominal conditions, for all realisations of disturbances w and v satisfying (2.37). The detection method described by the test (2.43) is said to be “conservative”, as it is a sufficient condition to assert whether the CPS is under attack. Indeed, if

$$|r(t)| > \bar{r}(t) \quad (2.44)$$

holds at any time instant $t \geq 0$, the CPS is guaranteed to be under attack. On the other hand, (2.43) may still be satisfied, even if any of the conditions in (2.42) do not hold.

Having described the centralised modelling framework for both the CPS and the attacking agent, let us now move on to the description of some specific attack strategies to be used in the analysis of the properties of the proposed distributed detection architectures.

2.2.3 Modelling attack strategies

We now introduce some classes of attacks. Throughout, we present their design characteristics, their properties, and the amount of resources required by the attacking agent to implement

them. Without the pretence of presenting an exhaustive list, we here introduce the following:

1. Bias-injection attacks;
2. Zero-dynamics attacks;
3. Replay attacks;
4. Covert attacks.

While the definitions we provide here are adapted from their description in [TSSJ15], for completeness we include references to other works that have analysed these attacks in detail.

Bias-injection attacks

When implementing bias injection attacks, malicious agents introduce a constant attack vector a_∞ into the communication channels between the controller/diagnoser unit and the plant $\mathcal{S}$. In its simplest form, it can be defined as

$$a(t) = \beta(t - T_a)a_\infty, \quad \forall t \geq T_a \quad (2.45)$$

with $T_a > 0$ the initial time of attack, $a_\infty \in \mathbb{R}^q$, and $\beta(\cdot)$ is an activation function to be defined. The impact of a bias-injection attack on the system is that of introducing a bias on the actuation or sensing channels, thus possibly disrupting the nominal operations of the overall system. We present the resources required for this class of attacks in Table 2.1.

We allow for the attacker to modulate the magnitude of its attacking signal by some function $\beta : \mathbb{R} \rightarrow \mathbb{R}$. Indeed, we consider that an attacking agent may assume the presence of a diagnosis

Table 2.1: Bias injection attacks – Attack resources

Model knowledge	Disclosure resources	Disruption resources
No model knowledge	No disclosure resources, i.e. no access to $u(t)$ nor $y(t)$ in real time	$B_a \neq 0$ and/or $D_a \neq 0$

unit monitoring the system, and it is well known in the fault diagnosis literature that abrupt faults may have a greater impact on diagnosis residuals [ZPP04]. Hence, defining function $\beta(\cdot)$:

$$\beta(t) = \begin{cases} 0 & \forall t < 0 \\ (1 - e^{-bt}) & \forall t \geq 0 \end{cases} \quad (2.46)$$

with $b > 0$, the possibility of detection to occur during the initial transient is reduced, for b and a_∞ in (2.45) sufficiently small. Note, clearly, that $\lim_{t \rightarrow \infty} \beta(t) = 1$, and as such

$$\lim_{t \rightarrow \infty} a(t) = \lim_{t \rightarrow \infty} \beta(t - T_a) a_\infty = a_\infty.$$

In order to assess the effect of this and the following classes of attacks on the residual error of the diagnoser $\mathcal{D}$, we exploit the principle of superposition. Specifically, let us decompose the residual r in

$$r = r^h + r^a, \quad (2.47)$$

where r^h is the nominal residual vector, while r^a is the component of the residual as affected by the attack. As expected, in nominal conditions $r^a(s) = 0, \forall s \leq T_a$. Therefore, the asymptotic effect of the attack on the residual in $\mathcal{D}$, by use of the final value theorem, and given stability of (2.39), can be computed as:

$$r_\infty^a = \lim_{t \rightarrow \infty} r^a = \lim_{s \rightarrow 0} (C_d(sI - A)^{-1} K_d + F_d) (C(sI - A)^{-1} B_a + D_a) a_\infty. \quad (2.48)$$

Remark 2.5. Note that, for the purpose of computing the effect of the attack on the residual error r , it is not necessary to consider its effect on the dynamic controller $\mathcal{C}$ in (2.38). Indeed, the controller's response to $a(t)$ affects both the dynamics of the anomaly detector in (2.39) and the plant dynamics in (2.36), thus not affecting the residual.

On the other hand, the attack's effect on the controller is fundamental to evaluate its overall effect on the state of the CPS. Indeed, again using the final value theorem, the component of

the state affected by the attack converges to:

$$x_\infty^a = \lim_{t \rightarrow \infty} x^a = \lim_{s \rightarrow 0} (I + G_{xy}(s))^{-1} G_{ya}(s) a_\infty$$

where

$$G_{ya}(s) = C(sI - A)^{-1} B_a + D_a \quad (2.49)$$

$$G_{xy}(s) = BC_c(sI - A_c)^{-1} B_c \quad (2.50)$$

are the attack-to-output transfer function of plant $\mathcal{P}$, which we assume to be stable for this result, and the output-to-state transfer function of controller $\mathcal{C}$, which is stable by construction. If the transfer function $G_{ya}(s)$ is marginally stable or unstable, a value x_∞^a may not exist. $\triangleleft$

From definitions of the detection test (2.43) and of the detection threshold $\bar{r}$, for a_∞ such that

$$|r_\infty^a| \leq \bar{r}(t), \quad \forall t \geq T_a \quad (2.51)$$

then it may not be detected by the diagnoser $\mathcal{D}$, if the activation function $\beta(\cdot)$ is chosen to be slow enough. However, given definition of stealthy attacks in Definition 2.7, it is necessary that the attack satisfy

$$r_\infty^a = 0 \quad (2.52)$$

for the malicious agent to *guarantee* that the effect of the attack is not detected.

Given the (clearly conservative) result above, the authors [TSSJ15] introduce the concept of α -stealthiness. A bias-injection attack is α -stealthy if it satisfies

$$|r_\infty^a| \leq \alpha \quad (2.53)$$

for some parameter $\alpha \geq 0$. However, in order to design such an attack, the malicious agent must have further knowledge of the dynamics of $\mathcal{S}$, which is not contemplated in our definition of bias-injection attacks.

Table 2.2: Zero-dynamics attacks – Attack resources

Model knowledge	Disclosure resources	Disruption resources
Full knowledge of dynamics (A, B, C) of $\mathcal{S}$	No disclosure resources	$B_a \neq 0$

Zero-dynamics attacks

Analysed by a number of works in literature, such as [PDB13, PDB15a], among others, zero-dynamics attacks exploit the *zero dynamics* of the physical system in open loop to achieve stealthiness. The resources required to perform such attacks are given in Table 2.2.

Indeed, zero-dynamics attacks can be seen as output-zeroing inputs, i.e. signals designed to achieve

$$y^a(t) = 0, \quad \forall t \geq T_a,$$

where we use superscript a to define the portion of y affected by the attack⁵. These can be seen as equivalent to unobservable inputs to the plant $\mathcal{S}$ given input matrix B_a [see Appendix A for a discussion on input observability]. Given that $y^a(t) = 0$, this class of attacks does not have any effect on the residual, and is therefore stealthy to $\mathcal{D}$.

The motivating idea behind zero-dynamics attacks is to design attacks which are stealthy to the diagnoser $\mathcal{D}$, while not having the necessary resources to perform attacks in closed loop, i.e. they can neither act directly on the measurements of the system, nor can they eavesdrop on the control signals sent from $\mathcal{C}$ to the plant $\mathcal{S}$. Indeed, they only depend on dynamics (2.36), and therefore only require knowledge of (A, B_a, C) of $\mathcal{S}$.

Therefore, considering that an attacker capable of performing a zero-dynamics attack has access to the disruption resources defined in Table 2.2, its effect on the state of the CPS can be computed as:

$$\begin{aligned} \dot{x}^a(t) &= Ax^a(t) + B_a a(t) \\ y^a(t) &= Cx^a(t) \end{aligned} \tag{2.54}$$

where, as previously, we use superscript a to indicate the component of any signal affected by

⁵Note that, given the attacker resources defined in Table 2.2, $\tilde{y} = y$.

an attack. Thus, supposing without loss of generality that B_a is full column rank, an attack $a(t)$ is *output zeroing* if it satisfies

$$a(t) = e^{\mu(t-T_a)} g, \quad (2.55)$$

where $\mu \in \mathbb{C}$ is a transmission zero of the LTI system (2.54), i.e. it is such that, given the Rosenbrock matrix

$$P^a(s) = \begin{bmatrix} sI - A & B_a \\ C & 0 \end{bmatrix}, \quad (2.56)$$

$P^a(\mu)$ loses rank. Furthermore, the vector $g \in \mathbb{R}^m$ satisfies

$$\begin{bmatrix} \mu I - A & B_a \\ C & 0 \end{bmatrix} \begin{bmatrix} x_0^a \\ g \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

with initial state $x_0^a = x^a(T_a)$. Given the definition of this attack strategy, if the transmission zero $\mu \in \mathbb{C}_-$, $\lim_{t \rightarrow \infty} a(t) = 0$ asymptotically. Hence, the most vulnerable cyber-physical systems are those that have zero-dynamics and whose transmission zeros are included in the open right hand complex plane.

Replay attacks

This attack strategy evolves over the course of two separate stages. Initially, starting at time T_r , the malicious agent performs a disclosure attack on the output measurements communicated between the plant and the controller/diagnoser and records the information. Subsequently, from time T_a , the attacker is capable of disrupting both the measurement and the actuation data, by transmitting different control signals u_a to $\mathcal{S}$ while *replaying* the recorded data back to $\mathcal{C}$ and $\mathcal{D}$ (hence its name).

Replay attacks have been studied by a number of works, including [MWS15, FT20], partially due to the fact that the Stuxnet virus employed a replay attack against a SCADA system to conceal its attack against a uranium enrichment facility in Iran [FMC11]. The resources necessary to carry out this attack can be found in Table 2.3.

Table 2.3: Replay attacks – Attack resources

Model knowledge	Disclosure resources	Disruption resources
No knowledge of the CPS dynamics is assumed	Disclosure attack possible on the output measurements y	$B_a \neq 0$ and $D_a \neq 0$

This class of attacks can be formalised as:

$$\begin{aligned} u_a(t) &= B_a a(t) \neq 0 \\ \tilde{y}(t) &= y(t - nT) - y(t) \end{aligned}, \quad \forall t \geq T_a \quad (2.57)$$

where $T \leq T_a - T_r$ is the period of the recorded data which is played back, while $n \in \mathbb{N}_0$ is a counter index, defined as

$$n \doteq \left\lceil \frac{t - T_a}{T} \right\rceil.$$

As pointed out in [MWS15], this attack is particularly insidious, as it does not require any information on the dynamics of the system, and, so long as the system is in steady state, cannot be easily detected: indeed, given that the malicious agent is capable of transmitting information $y(t), t \in [T_r, T_a]$, its trajectory presents the dynamic signature of the CPS, while also embedding the statistical properties of the noise.

If, however, the attack is not carried out when $\mathcal{S}$ is in steady state, the condition for stealthiness is not trivial, as shown in Chapter 3 when analysing the performance of the distributed diagnoser $\mathcal{D}_i$ against replay attacks.

Covert attacks

This class of attacks, the most sophisticated among those analysed in this introductory chapter, was analysed in [Smi11, Smi15, BRBP20, BGBP19]. These attacks require the most amount of knowledge on the part of the attacker: indeed, the malicious agent must have knowledge of the dynamics of the system *a priori*, while online it must be capable of disrupting both the actuation and sensing channels of $\mathcal{S}$. The resources required to carry out a covert attack are found in Table 2.4.

When implementing an attack in this class, the malicious agent is capable of injecting an input signal to the plant, while simultaneously removing its effect from the measurement data

Table 2.4: Covert attacks – Attack resources

Model knowledge	Disclosure resources	Disruption resources
(A, B, C) known exactly	Availability of u and y	Ability to overwrite u and $\tilde{y}$

transmitted by $\mathcal{S}$ *exactly*. It does so by including a closed-loop feedback structure within the attack. Formally, this class of attacks may be defined as the combination of the following LTI dynamical system:

$$\begin{cases} \dot{x}_a = A^a x_a + B^a u_a \\ y_a = C^a x_a \end{cases} \quad (2.58)$$

with $x_a \in \mathbb{R}^n, y_a \in \mathbb{R}^p, u_a \in \mathbb{R}^{m_a}$ the attacker's internal state, input and output, with appropriately defined matrices, and initial condition $x_a(T_a)$, and the attack vector $a(t)$ satisfying the following conditions:

$$\begin{aligned} \tilde{u}(t) &= u(t) + B_a a(t) = u(t) + u_a(t), \\ \tilde{y}(t) &= y(t) + D_a a(t) = y(t) - y_a(t), \end{aligned} \quad \forall t \geq T_a, \quad (2.59)$$

recalling the definition of B_a and D_a in (2.40).

In the following Lemma, adapted from [Smi15], where covert attacks are formulated in the frequency domain, we give conditions for covert attack stealthiness. Note that the condition for stealthiness here is similar to those given in [BRBP20] for the distributed case.

Lemma 2.1. Consider a covert attack as in (2.58)-(2.59) against a CPS with dynamics (2.36). Suppose that the malicious agent has perfect knowledge of the CPS dynamics, and that:

$$(A^a, B^a, C^a) = (A, B, C).$$

If $x_a(T_a) = 0$, the covert attack will be stealthy to the diagnosis module $\mathcal{D}$ as defined in (2.39).

□

Remark 2.6. Note that there is a strong relationship between the behaviour of covert and replay attacks. Indeed, apart from the time instances in the set $\mathcal{T} \doteq \{t \in \mathbb{R} | t = T_a + nT, n \in \mathbb{Z}\}$,

$\mathbb{N}_0\}$, the dynamics of the replay attack are the same as (2.58), with

$$u_a = u(t - nT), \quad x_a(T_a) = x(T_a - T). \quad (2.60)$$

Given the periodic nature of the attack, this does not hold for $t \in \mathcal{T}$, where, depending on the difference between $x(T_{a-})$ and $x(T_a - T)$, there may be a discontinuity. $\triangleleft$

2.3 Conclusions

In this chapter we presented the necessary background to model DC microgrids, the main application of this thesis. Specifically, we gave the physical dynamics of each DGU, building up to their state-space description, showing how the switching nature of the control input can be accounted for.

We also summarised the modelling framework of cyber-attacks in cyber-physical systems, from a centralised perspective. Furthermore, we introduced the modelling techniques for a number of classes of attacks, as proposed in [TSSJ15]. Relying on the description of the *attack space* in [TSSJ15], shown in Figure 2.3, we have presented each class of stealthy attack focusing on the resources required, in terms of *a priori* knowledge, and disclosure and disruption resources.

Chapter 3

Distributed cyber-attack detection

As discussed in the previous chapters, the problem of cyber-attacks in networked control systems, such as future energy distribution networks, poses a severe threat to their operation, putting the safety of personnel, users and infrastructure at risk. As such, detection methods must be designed to detect attacks in a timely manner.

Given the nature of large-scale systems, their size and complexity are such that centralised diagnosis techniques either are not feasible, or they require costly investments to embed the necessary computational and communication resources. As such, distributed architectures for control and fault detection and isolation of LSSs have attracted large amounts of attention in the literature (e.g. [BKLS16, STSJ11, TSSJ14, DMK16, BFK⁺17, BRFTP19], to name a few).

The main objective of this chapter is to introduce and analyse a cyber-attack detection technique for large-scale systems. A distributed detection architecture is chosen, enabling scalable implementation of the detection tool in partitioned LSSs with non-overlapping decomposition [Lun92].

Although the results here have been generalised to a class of large-scale interconnected systems, the leading motivation behind its design was its application to DC microgrids. Indeed, the development of the proposed architecture is a direct derivation of the set of issues posed by this application. Therefore, once the characteristics of the detection architecture are given for general systems, we turn our attention to DC microgrids as a case study, showing through realistic simulations the appropriateness of the monitoring strategy to this application.

In the remainder of this chapter, we start by formally introducing the problem of distributed cyber-attack detection within the communication network linking different subsystems. Following this, the distributed detection architecture is presented algorithmically, without providing its technical details. The proposed architecture relies on two parallel modules working simultaneously to detect attacks: one is based on detection through a Luenberger observer, while the other exploits a bank of unknown-input observers. Before introducing the two modules, we give structural limitations of distributed diagnosers when the information available for the design of the diagnoser is restricted to only including local subsystem information: this may be due to some design constraints, such as the need for privacy, or limitations in available computational power.

Following this, the characteristics of each of the modules are presented, together with their joint properties within the distributed architecture. Specifically, we show how, by using different information structures, the combination of the two modules is capable of improving the performance of the diagnoser. Finally, the application of this architecture to DC microgrids is presented in detail, offering both theoretical and simulation-based results.

The work presented in this chapter draws from results shown in the publications:

- [BGFTP17] F. Boem, A. J. Gallo, G. Ferrari-Trecate, T. Parisini, “A distributed attack detection method for multi-agent systems governed by consensus-based control”, 56th Annual Conference on Decision and Control (CDC), pp. 5961–5966, 2017
- [GTN⁺18] A. J. Gallo, M. S. Turan, P. Nahata, F. Boem, T. Parisini, G. Ferrari-Trecate, “Distributed cyber-attack detection in the secondary control of DC microgrids”, European Control Conference (ECC), pp. 344–349, 2018.
- [GTB⁺20] A. J. Gallo, M. S. Turan, F. Boem, T. Parisini, G. Ferrari-Trecate, “A Distributed Cyber-attack Detection Scheme with application to DC Microgrids”, IEEE Transactions on Automatic Control, vol. 65, n. 9, pp. 3800–3815, 2020.

- [GBP20] A. J. Gallo, A. Barboni, T. Parisini, “On detectability of cyber-attacks for large-scale interconnected systems”, 21st IFAC World Congress, 2020, *To appear*

3.1 Problem formulation

Having introduced the problem of cyber-attack detection in Chapter 1, and given a general overview of its centralised solution in Chapter 2, we now formally introduce the scenario considered when dealing with *distributed cyber-attack detection*. After addressing the modelling framework used to describe large-scale systems as the interconnection of a number of subsystems, we introduce the source of possible cyber-attacks in distributed networked control architectures. Thus, we define the problem of *distributed* and *scalable* cyber-attack detection.

3.1.1 Modelling large-scale systems

Similarly to the modelling framework presented in Chapter 2.1 for DC microgrids as networks of interconnected DGUs, but differently from the centralised model of cyber-physical systems presented in Chapter 2.2, we now model LSSs as networks of linear-interconnected subsystems, exploiting a non-overlapping partition method, as described in [Lun92]. We show this pictorially in Figure 3.1. Specifically, we consider a set of N subsystems $\mathcal{S}_i$, where $i \in \mathcal{N} \doteq \{1, \dots, N\}$ is used as an index. Each subsystem is interconnected to a set $\mathcal{N}_i$ of neighbouring subsystems, which dynamically interact with the state x_i , i.e., $\mathcal{N}_i \doteq \{j \in \mathcal{N} | \partial x_i / \partial x_j \neq 0\}$, with $N_i \doteq |\mathcal{N}_i|$. The dynamics of $\mathcal{S}_i$ are defined as:

$$\mathcal{S}_i : \begin{cases} \dot{x}_i = A_{ii}x_i + B_i u_i + E_i d_i + \xi_i + w_i \\ y_i = C_i x_i + v_i \end{cases} \quad (3.1)$$

where $x_i \in \mathbb{R}^{n_i}$, $u_i \in \mathbb{R}^{m_i}$, $d_i \in \mathbb{R}^{q_i}$, $y_i \in \mathbb{R}^{p_i}$ are respectively the state, control input, exogenous input, and output of $\mathcal{S}_i$; moreover, $\xi_i \in \mathbb{R}^{n_i}$ models the *physical coupling* between subsystems,

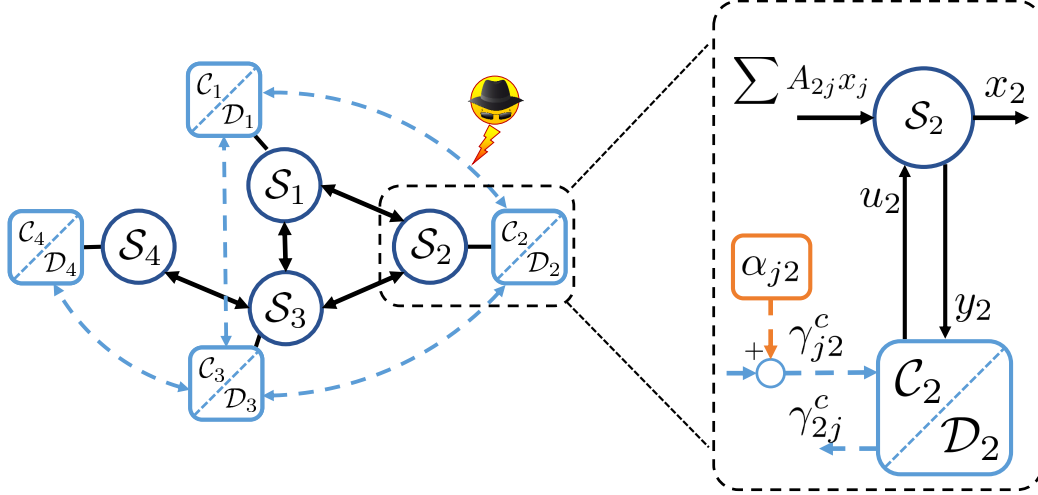


Figure 3.1: Representation of a LSS as a network of $N = 4$ subsystems. [On the left] Networked subsystems. Each subsystem is equipped with a controller $\mathcal{C}_i$ and a diagnoser $\mathcal{D}_i$. The physical coupling between subsystems is represented by the black arrows, while the communication network amongst controllers is shown through the blue arrows. [On the right] Subsystem $\mathcal{S}_2$ with its control and monitoring layer. The same color convention as the left is maintained.

with

$$\xi_i \doteq \sum_{j \in \mathcal{N}_i} A_{ij} x_j; \quad (3.2)$$

vectors $w_i \in \mathbb{R}^{n_i}$, $v_i \in \mathbb{R}^{p_i}$ represent state and output disturbances, both assumed to be bounded. In (3.1) all matrices are of the appropriate dimensions.

Assumption 3.1. *For all $\mathcal{S}_i$, $i \in \mathcal{N}$, time-varying signals $w_i(t)$ and $v_i(t)$ are unknown-but-bounded, i.e.:*

$$|w_i(k)| \leq \bar{w}_i, \quad |v_i(k)| \leq \bar{v}_i, \quad \forall k \geq 0 \quad (3.3)$$

where bounds $\bar{w}_i < \infty$, $\bar{v}_i < \infty$ are known a priori. $\triangleleft$

Assumption 3.2. *For all subsystems $\mathcal{S}_i$, $i \in \mathcal{N}$:*

- (A_{ii}, B_i) is stabilizable;
 - (A_{ii}, C_i) is detectable;
 - $\text{rank } C_i = p_i \leq n_i$.
- $\triangleleft$

Assumption 3.3. For all $i \in \mathcal{N}$, matrices C_i are such that

$$\ker(C_i) \subseteq \ker A_{ji}, \quad (3.4)$$

for all $i \in \mathcal{N}_j$. ◁

Remark 3.1. Assumption 3.3 is a necessary condition for Theorem 3.1 in Section 3.2. Furthermore, it eases presentation of Section 3.4: indeed, in Section 3.4.3 we introduce a method to relax it. ◁

Remark 3.2. It is worth noting that the dynamics of the overall large-scale system in (2.36) can be reconstructed starting from the dynamics of its partitioned subsystems (3.1). Indeed, defining $s \doteq \text{col}_{i \in \mathcal{N}}(s_i)$ for all signal variables s_i in (3.1), the dynamics of the LSS can be written as

$$\mathcal{S} : \begin{cases} \dot{x} = Ax + Bu + Ed + w \\ y = Cx + v \end{cases} \quad (3.5)$$

with $A \doteq [A_{ij}]$, $B \doteq \text{diag}_{i \in \mathcal{N}}(B_i)$, $C \doteq \text{diag}_{i \in \mathcal{N}}(C_i)$, $E \doteq \text{diag}_{i \in \mathcal{N}}(E_i)$. Clearly, given Assumption 3.2, (A, B) is stabilizable and (A, C) is detectable. ◁

3.1.2 Distributed control strategy

We suppose that each subsystem $\mathcal{S}_i$ is equipped with a *distributed* controller $\mathcal{C}_i$. The distributed controller is designed such that global stability of the LSS and global objectives requiring coordination amongst subsystems (i.e. consensus) are achieved. To coordinate multiple controllers, a communication network is considered to be integrated in the system (see [CBCZ16] for a discussion on the need for the integration of communication resources within the voltage regulation problem). We define the control input as:

$$u_i = \kappa_i(y_i, \gamma_{ji, j \in \mathcal{N}_i}^c), \quad (3.6)$$

where $\kappa_i(\cdot, \cdot)$ is an operator representing the distributed control law applied by $\mathcal{C}_i$. Note that we do not restrict κ_i to any specific class of control laws; rather, we identify it as depending on both

local measurement outputs y_i , and signals $\gamma_{ji,j \in \mathcal{N}_i}^c$, formally defined in (3.7). Indeed, $\kappa_i(\cdot, \cdot)$ can also be used to model the plug-and-play distributed control protocol defined in Chapter 2.1.2 for DC microgrids. The signal γ_{ji}^c is used to represent the information the controller $\mathcal{C}_i$ receives from the controller $\mathcal{C}_j$ associated to its neighbour $\mathcal{S}_j$, defined in the following.

Assumption 3.4. Define $\mathcal{G}^p \doteq \{\mathcal{N}, \mathcal{E}^p\}$ the graph induced by the physical coupling between subsystems, where the pair $(i, j) \in \mathcal{E}^p \iff A_{ij} \neq 0$. The graph induced by the communication network between controllers $\mathcal{G}^c \doteq \{\mathcal{N}, \mathcal{E}^c\}$ has the same topology as $\mathcal{G}^p$, i.e. $\mathcal{E}^c \equiv \mathcal{E}^p$. $\triangleleft$

Assumption 3.5. We consider an ideal communication network, i.e. all information is transmitted instantaneously without delay, packet drop-outs, or noise. $\triangleleft$

3.1.3 Cyber-attacks

As argued in Chapters 1 and 2, the introduction of a communication network within the control loop exposes the system to the possibility of cyber attacks. In the scenario we consider here, because of the completion of the distributed control loop *between* controllers $\mathcal{C}_i$ through the integration of a communication network, this exchanged information may be exposed to interference by malicious agents. We therefore formally define the variable γ_{ji}^c received by $\mathcal{C}_i$ from $\mathcal{C}_j$ to include a possible attack signal:

$$\gamma_{ji}^c(t) = C_j x_j(t) + v_j(t) + \beta_{ji}(t - T_a^{ji}) \alpha_{ji}(k) \quad (3.7)$$

where $\beta_{ji}(\cdot)$ is an activation function, α_{ji} is the attack signal and $T_a^{ji} > 0$ is the activation time. Function $\beta_{ji} : \mathbb{R} \rightarrow \mathbb{R}$ is such that $\beta_{ji}(s) = 0, \forall s < 0$. The activation function is commonly defined as [ZPP02]:

$$\beta_{ji}(s) = \begin{cases} 0, & \text{for } s < 0 \\ (1 - e^{-h_{ji}s}), & \text{for } s \geq 0 \end{cases} \quad (3.8)$$

with $h_{ji} > 0$. Note that for small values of h_{ji} the onset of the cyber-attack is slow, i.e. it can be seen as an *incipient* attack; on the other hand, as $h_{ji} \rightarrow \infty$, $\beta_{ji}(\cdot)$ approaches a step function, describing an *abrupt* attack, borrowing from the terminology used in fault detection and isolation literature.

Assumption 3.6. *Each communication channel $(j, i), \forall j \in \mathcal{N}_i, \forall i \in \mathcal{N}$ is affected by at most one attack, and $T_a^{ji} > 0$.* $\triangleleft$

Remark 3.3. Although limiting the scope of all possible attacks, Assumption 3.6 does not prevent multiple attacks in the communication network from coordinating, and is a common assumption in literature. $\triangleleft$

Remark 3.4. Throughout this chapter, we implicitly consider that each communication line between subsystems can be altered by an attack independently of the others, which is consistent with wired communication networks. This however can be extended to wireless communication networks by considering that all signals broadcast by a subsystem's controller, or all those received, are altered at once. $\triangleleft$

The attack function $\alpha_{ji}(t)$ is designed by the malicious agent to achieve some objective, and is unknown to $\mathcal{D}_i$. Recall a few examples of the specific strategies adopted by an attacker are given in Chapter 2.

Assumption 3.7. *For all $t \geq 0$, the local signals $u_i(t)$ and $y_i(t)$ are known exactly to $\mathcal{C}_i$ and $\mathcal{D}_i$.* $\triangleleft$

Remark 3.5. We stress that, differently from the majority of the literature, and from the centralised case in Chapter 2, we consider that the communication network *between* controllers is exposed to attacks. As such, we do not consider any attack on the local output measurements y_i , nor on the control input signal u_i , as given in Assumption 3.7. This is consistent with the considered application, i.e. that of power-electronics-enabled distribution networks, as mentioned in Chapter 1.2 for the case of DC microgrids, where the controllers, actuators, plant and sensors are all colocated. $\triangleleft$

3.1.4 Distributed cyber-attack detection

We are now able to formally define the problem to be solved:

Problem 3.1. Given subsystem $\mathcal{S}_i$, with dynamics as in (3.1), and distributed controller $\mathcal{C}_i$, design a *distributed* diagnosis module $\mathcal{D}_i$ capable of verifying that the null condition:

$$\mathcal{H}_i^0 \doteq \{t < T_a^{ji}, \forall j \in \mathcal{N}_i\} \quad (3.9)$$

holds, i.e. $\alpha_{ji}(\tau) = 0$ for all γ_{ji}^c , $j \in \mathcal{N}_i$, and all $\tau = [0, t)$. ◀

To avoid confusion, we explicitly clarify the meaning we give to *distributed* monitoring systems:

Definition 3.1 (Distributed module). *A module (be it for control or monitoring) designed for $\mathcal{S}_i$ is distributed if its design and online operations require information from $\mathcal{S}_k, k \in \mathcal{N}_i \cup \{i\}$, not from $\mathcal{S}$ as a whole. during operation, it relies on input and output information from $\mathcal{S}_k, k \in \mathcal{N}_i \cup \{i\}$, and not from $\mathcal{S}$ as a whole.* ◁

Another property that is beneficial for subsystem modules within the context of large-scale systems is that of being “plug-and-play” [Sto09, RFFT13]:

Definition 3.2 (Plug-and-play [RFFT13]). *A monitoring or control module for a subsystem $\mathcal{S}_i$ within an LSS is plug-and-play if it satisfies the following characteristics:*

i) *the design of a single module depends at most on information about subsystems $\mathcal{S}_k, k \in \mathcal{N}_i \cup \{i\}$;*

ii) *when a subsystem $\mathcal{S}_i$ joins or leaves an LSS, there is an automatic procedure to*

- *assess whether the operation spoils the stability of the LSS*
- *tune at most the subsystem and its successors, i.e. those subsystems $\mathcal{S}_j$ such that $i \in \mathcal{N}_j$.* ◁

Note that, because we are interested here in the design of a monitoring architecture for $\mathcal{S}$, its operation does not spoil the stability of the LSS as a whole. As such, we are not interested in the feasibility of a plug-in event.

The main objective of this chapter is to present the design and properties of distributed cyber-attack detection module $\mathcal{D}_i$, showing how its design can be plug-and-play.

3.2 Structural detectability for distributed cyber-attack diagnosers

Let us first introduce some structural detectability properties of distributed cyber-attack detectors designed with *reduced-information sets*. Formally defined below, modules based on reduced-information sets rely only on local subsystem information for the design of the detection modules, while exploiting possibly attacked information from neighbouring subsystems for online operations. Although computationally efficient and capable of preserving privacy among subsystems¹, this class of detectors are susceptible to *structurally stealthy attacks*² [Definition 3.6].

Let us consider the dynamics of $\mathcal{S}_i$ in (3.1). For the purpose of this section³, we suppose $\bar{w}_i = 0$ and $\bar{v}_i = 0$ for all $i \in \mathcal{N}$. Indeed, we here are focused on the detectability properties of a distributed diagnosis module structural to its design, and therefore are not influenced by the additive noise terms w_i, v_i .

Definition 3.3 (Strong observability [TSH12, Def. 7.15]). *A system Σ with dynamics described by*

$$\Sigma : \begin{cases} \dot{\xi} = A\xi + Bu \\ \zeta = C\xi + Du \end{cases}$$

is called strongly observable if for all initial conditions $\xi_0 \in \mathbb{R}^n$ and for any input function u the following holds: $\zeta(t, \xi_0) = 0$ for all $t \geq 0$ implies $\xi_0 = 0$. $\triangleleft$

Remark 3.6. The definition of strong observability introduced in Definition 3.3, taken from [TSH12], introduces the idea that for a strongly observable system there are no inputs $u(t) \neq 0$ such that the output $\zeta(t)$ is zero for all $t \geq 0$ with $\xi_0 \neq 0$. This differs from the property of *observability* of a system, for which it is important to evaluate the *distinguishability* of the output from the initial conditions, given a known input; indeed, for strong observability, the

¹Here “privacy” is used rather loosely. However, given that for this class of distributed modules no model knowledge of the LSS apart from that of the local subsystem itself is needed, no subsystem must share its model information with any of its neighbours.

²The results in this section were presented in [GBP20].

³Note that, in the following sections of this chapter, $\bar{w}_i \neq 0$ and $\bar{v}_i \neq 0$ are considered.

input u may be considered to be unknown. We refer the interested reader to [TSH12, Ch. 7.5] for further detail on strongly observable systems and their properties. $\triangleleft$

Assumption 3.8. *For subsystem $\mathcal{S}_i$ with dynamics as in (3.1), the system described by the tuple $(A_{ii}, \check{A}_{ij}, C_i)$ is strongly observable for each $j \in \mathcal{N}_i$, where $\check{A}_{ij}$ is a matrix of independent columns forming a basis of $\text{Im}A_{ij}$.* $\triangleleft$

In this section, we are interested in analysis of those properties which arise following the design of distributed diagnosers and the partition of the LSS into subsystems. Specifically, we investigate the properties of *reduced-information distributed diagnosers*.

Definition 3.4 (Information set). *The information set $\mathcal{I}_i$ available for the design of $\mathcal{D}_i$ is composed of the offline model knowledge available to the diagnoser and the online signals it exploits for detection. Specifically, the information set is defined as*

$$\mathcal{I}_i \doteq \{\mathcal{M}_i, \gamma_i, v_i\}, \quad (3.10)$$

where $\mathcal{M}_i$ is a tuple containing a partition of the dynamics of the LSS available to $\mathcal{D}_i$ a priori, while γ_i and v_i are the output and input signals of the LSS available to the diagnoser during online operation. $\triangleleft$

Definition 3.5 (Reduced-information distributed detection architecture). *A distributed detection architecture is said to be reduced information when the information set used for design is*

$$\mathcal{I}_i \doteq \{(A_i, B_i, A_{ij, j \in \mathcal{N}_i}, C_i), [y_i^\top, \gamma_{ji, j \in \mathcal{N}_i}^{c\top}]^\top, u_i\}. \quad (3.11)$$

If the diagnoser $\mathcal{D}_i$, $i \in \mathcal{N}$ is designed based on this set of information, it is said to be “reduced information” itself. $\triangleleft$

Because the diagnoser lacks direct access to x_j , $j \in \mathcal{N}_i$, it must construct a residual generator to detect attacks on the received information γ_{ji}^c . To be valid for attack detection, the residual r_i must satisfy

$$\lim_{t \rightarrow \infty} r_i(t) = 0 \quad (3.12)$$

when not under attack, while being sensitive to attacks, i.e. $r_i \neq 0$ for $\alpha_{ji} \neq 0$. Given linearity of the dynamics of $\mathcal{S}_i$, the residual can be rewritten as $r_i = r_i^h + r_i^a$, where r_i^h is the *healthy*⁴, or nominal, component of the residual, while r_i^a is the component of r_i directly affected by the attack. Similarly to the definition of stealthy attacks in Chapter 2, we introduce the concept of *structurally stealthy attacks*.

Definition 3.6 (Structurally stealthy attacks). *An attack $\alpha_{ji}(t)$ on the transmitted signal $\gamma_{ji}^c(t)$ is said to be structurally stealthy to $\mathcal{D}_i$, given information set $\mathcal{I}_i$, if $\alpha_{ji}(t) \neq 0$ in (3.7) leads to $r_i^a(t) = 0$ for all $t \geq 0$.* $\triangleleft$

We stress that the distinction between *stealthy* and *structurally stealthy* attacks is an important, albeit subtle, one. Indeed, stealthy attacks in Definition 2.7 are defined as a property of attacks, that exploit their own resources to construct an attack that is stealthy. On the other hand, for structurally stealthy attacks defined in Definition 3.6, it is a property of attacks *in relation* to the information set $\mathcal{I}_i$. The following example is given to highlight this conceptual difference.

Example 3.1. We start by noting that the residual r_i can be seen as dependent on $\mathcal{I}_i$. Indeed, suppose $\mathcal{I}_i$ includes direct knowledge of y_j for all $j \in \mathcal{N}_i$. Then an appropriate residual to detect attacks $\alpha_{ji}(t) \neq 0$ in the communicated signal (3.7) could be defined as:

$$r_i(t) = \text{col}_{j \in \mathcal{N}_i} (y_j(t) - \gamma_{ji}^c(t)).$$

Let us define the initial instant of attack

$$T_a \doteq \min_{j \in \mathcal{N}_i} T_a^{ji}.$$

Under nominal conditions, i.e. for $t \leq T_a$, the information received by $\mathcal{C}_i$ from $\mathcal{C}_j$ is exactly the output measurement: $\gamma_{ji}^c(t) = y_j(t)$, for all $j \in \mathcal{N}_i$. Thus,

$$r_i(t) = 0, \forall t \leq T_a.$$

⁴Here, and throughout, we use *healthy* intending the component of a variable the behaviour of which coincides with its *nominal* behaviour [Definition 2.3].

On the other hand, when under attack $\alpha_{ji} \neq 0, j \in \widehat{\mathcal{N}}_i(t) \doteq \{j \in \mathcal{N}_i | t \geq T_a^{ji}\}$ and

$$r_i = \text{col}_{j \in \widehat{\mathcal{N}}_i(t)} (\alpha_{ji}) \neq 0.$$

As such, no attack $\alpha_{ji} \neq 0$ would be structurally stealthy to any $\mathcal{D}_i$ for which $y_j \in \mathcal{I}_i$. ∇

Remark 3.7. Note that Example 3.1 highlights that, under certain conditions, it is possible to have diagnosers without any structurally stealthy attacks. Furthermore, it illustrates how, by changing the information set $\mathcal{I}_i$, the class of attacks structurally stealthy to $\mathcal{D}_i$ changes. $\triangleleft$

For the scenario we are considering here, i.e. the solution of Problem 3.1, given the definition of the reduced-information sets used in design of the diagnosers, the neighbouring output measurements y_j are not included in $\mathcal{I}_i$. As such, an appropriate residual generator cannot take the form presented in Example 3.1.

For the main result of this section, we consider that a single communication link (j, i) is under attack from some time $T_a = T_a^{ji} > 0$. In Corollary 3.1, we introduce a similar result for $|\widehat{\mathcal{N}}_i(t)| > 1$, with $\widehat{\mathcal{N}}_i(t) \doteq \{j \in \mathcal{N}_i | t \geq T_a^{ji}\}$ defining the subset of neighbours whose communications are subject to attacks.

Theorem 3.1. Consider a subsystem with dynamics as in (3.1) equipped with a reduced-information diagnoser $\mathcal{D}_i$, and that Assumption 3.3 holds. Suppose that, from some time $T_a > 0$, $\alpha_{ji}(t) \neq 0$, for some unknown $j \in \mathcal{N}_i$. Such an attack is structurally stealthy if and only if

$$C_j^\dagger \alpha_{ji}(t) \in \ker A_{ij} \tag{3.13}$$

holds for all $t \geq T_a$. $\square$

Proof. **[Sufficiency]:** Recall that, in order for the diagnoser to detect an attack, the residual r_i must be such that, given an attack $\alpha_{ji} \neq 0$, if $\gamma_{ji}^c \neq y_j$, then $r_i \not\rightarrow 0$. As the control and exogenous inputs are either known or measured by $\mathcal{D}_i$, and given that they cannot be altered maliciously [Assumption 3.7], their effect on the state trajectory are disregarded. Given that the state and model of $\mathcal{S}_j$ are unknown to the local diagnostic module $\mathcal{D}_i$, x_j must be *statically*

estimated to achieve $r_i \rightarrow 0$ in nominal conditions, while still being sensitive to attacks. Indeed, although an unknown-input observer⁵ could be used to estimate x_i , and thus to achieve $r_i \rightarrow 0$ irrespective of x_j , this would, by construction, not be sensitive to any attack $\alpha_{ji} \neq 0$. It is therefore an inappropriate choice. We now proceed with the definition of the residual generator exploiting a static estimate of x_j .

Disregarding the effect of initial conditions $x_i(0)$ and of other subsystems $\mathcal{S}_k, k \in \mathcal{N}_i \setminus \{j\}$, consider the map from x_j to y_i :

$$y_i(t) = C_i \int_0^t e^{A_{ii}(t-\tau)} A_{ij} x_j(\tau) d\tau. \quad (3.14)$$

Hence, a residual generator of the form

$$\begin{aligned} r_i(t) &= y_i - \hat{y}_i \\ &= y_i - C_i \int_0^t e^{A_{ii}(t-\tau)} A_{ij} \hat{x}_j(\tau) d\tau \end{aligned} \quad (3.15)$$

may be considered, where $\hat{y}_i$ is the local output estimate and $\hat{x}_j$ is the estimate of x_j as computed by $\mathcal{D}_i$. An appropriate static estimate of x_j is

$$\hat{x}_j \doteq C_j^\dagger \gamma_{ji}^c, \quad (3.16)$$

where $C_j^\dagger$ is unique, given Assumption 3.2. Since $\text{rank } C_j = p_j \leq n_j$, this method may introduce an error $e_{ji} \doteq x_j - C_j^\dagger \gamma_{ji}^c \neq 0$. However, this does not affect the residual generator, as in nominal conditions, i.e. for $\alpha_{ji} = 0$, any error is such that

$$e_{ji} = (I - C_j^\dagger C_j) x_j \in \ker C_j \quad (3.17)$$

by construction, while $e_{ji} \in \ker A_{ij}$ holds given Assumption 3.3. Thus from definition of residual r_i in (3.15), $r_i = 0$ even if $e_{ij} \neq 0$.

To prove sufficiency, suppose that α_{ji} is designed such that (3.13) holds. From linearity, we

⁵Note that some technical assumptions must be fulfilled by the dynamics of the subsystem $\mathcal{S}_i$ for an unknown-input observer to be designed – see Appendix A for details.

can decompose the estimated output $\hat{y}_i = \hat{y}_i^h + \hat{y}_i^a$, where $\hat{y}_i^h$ is the *healthy* component of the output estimate. On the other hand, $\hat{y}_i^a$ is the *attacked* component of the output estimate, and it is such that:

$$\hat{y}_i^a = C_i \int_{T_a}^t e^{A_{ii}(t-\tau)} A_{ij} C_j^\dagger \alpha_{ji}(\tau) d\tau = 0. \quad (3.18)$$

Hence, given that after an attack

$$r_i \rightarrow \hat{y}_i^a,$$

for $t \geq T_a$, the residual satisfies $r_i(t) = 0$, as $\hat{y}_i^a(t) = 0$, for all $t \geq T_a$, if condition (3.13) holds. Thus, the attack is structurally stealthy.

[Necessity]: We prove necessity by contradiction. Let us suppose that $C_j^\dagger \alpha_{ji} \notin \ker A_{ij}$, but that the attack is indeed structurally stealthy, i.e. a diagnoser $\mathcal{D}_i$ with reduced-information set $\mathcal{I}_i$ as defined in (3.11) is not capable of detecting it. Decompose $C_j^\dagger \alpha_{ji} = C_j^\dagger \bar{\alpha}_{ji} + C_j^\dagger \alpha_{ji}^\perp$, with $C_j^\dagger \bar{\alpha}_{ji} \in \ker A_{ij}$ and $C_j^\dagger \alpha_{ji}^\perp \in (\ker A_{ij})^\perp$, and such that

$$C_j^\dagger \alpha_{ji} = C_j^\dagger \bar{\alpha}_{ji} + C_j^\dagger \alpha_{ji}^\perp.$$

The residual generator defined in (3.15) approaches

$$r_i(t) \rightarrow C_i \int_{T_a}^t e^{A_{ii}(t-\tau)} A_{ij} C_j^\dagger \alpha_{ji}^\perp d\tau \neq 0 \quad (3.19)$$

as $t \rightarrow \infty$, where the last equality is given by the fact that $C_j^\dagger$ is full column rank by construction and (3.13) is supposed not to hold, in turn implying $C_j^\dagger \alpha_{ji}^\perp \neq 0$. Furthermore, given Assumption 3.8, there are no zero dynamics between A_{ij} and C_i : hence r_i is sensitive to any attack. Thus by contradiction we prove the claim. $\blacksquare$

Theorem 3.1 can be easily extended for attacks on multiple communication lines, i.e. for $|\widehat{\mathcal{N}}_i| > 1$:

Corollary 3.1. Consider a subsystem with dynamics as in (3.1) equipped with a reduced-information diagnoser $\mathcal{D}_i$. Suppose that, from time $T_a > 0$, $\alpha_{ji}(t) \neq 0$, for some set of unknown

$j \in \widehat{\mathcal{N}}_i \subseteq \mathcal{N}_i$, with $\widehat{\mathcal{N}}_i$ such that

$$\sum_{j \in \widehat{\mathcal{N}}_i} n_j > n_i.$$

Such attacks are structurally stealthy if and only if

$$\bar{C}_{ji}^\dagger \bar{\alpha}_{ji}(t) \in \ker \bar{A}_{ij}, \quad \forall t \geq T_a, \quad (3.20)$$

with $\bar{C}_{ji} \doteq \text{diag}(C_j)_{j \in \widehat{\mathcal{N}}_i}$, $\bar{\alpha}_{ji}(t) \doteq \text{col}(\alpha_{ji}(t))_{j \in \widehat{\mathcal{N}}_i}$, and $\bar{A}_{ij} \doteq \text{row}(A_{ij})_{j \in \widehat{\mathcal{N}}_i}$. $\square$

We note that the result presented in Theorem 3.1 shows some similarities to the class of zero-dynamics attacks in literature, which have been summarised in Chapter 2 – even though these contradict Assumption 3.8. Indeed, the physical coupling between subsystems can be seen as an exogenous input to $\mathcal{S}_i$ which are not fully input observable, and γ_{ji}^c can be seen as *measurements* of those inputs. As such, the class of attacks identified through Theorem 3.1 can be seen as a subset of the class of zero-dynamics attack, as a reduced-information detector $\mathcal{D}_i$ lacks the redundancy to be able to detect these attacks.

A rather significant difference with zero-dynamics attacks as defined in Chapter 2, however, is that an attacker can design a structurally stealthy attack to a reduced-information monitor $\mathcal{D}_i$ by exploiting knowledge of A_{ij} , rather than the full knowledge of dynamics of $\mathcal{S}_i$, and specifically knowledge of its nullspace. Furthermore, these attacks, as all of the attacks considered in this chapter, act upon the communicated signals γ_{ji}^c , and not on the subsystems directly. As such, their effect on $\mathcal{S}_i$ occurs through the computation of the input u_i by $\mathcal{C}_i$.

Although the class of attacks defined in Theorem 3.1 relies on a smaller amount of model information than the standard zero-dynamics attacks described in literature, it is well known that accurate knowledge of interconnection matrices A_{ij} is generally hard to obtain. As such, in the following we present a subclass of structurally stealthy attacks to reduced-information $\mathcal{D}_i$, by looking at the structural matrix $[A_{ij}]$, i.e. the matrix describing the so called “zero-pattern” of A_{ij} .

Definition 3.7 (Structural matrices). *For a given matrix M , $[M]$ is used to define its corresponding structural matrix. Each ij -th entry of $[M]$ is written as $[m]_{ij}$ and is related to the*

corresponding element m_{ij} of M through the following relation:

$$[m]_{ij} \doteq \begin{cases} 0, & \text{if } m_{ij} = 0; \\ *, & \text{otherwise.} \end{cases} \quad (3.21)$$

In (3.21), $*$ represents a free parameter. $\triangleleft$

Corollary 3.2. Suppose $\text{rank}[A_{ij}] < n_j$. It is sufficient that an attack $\alpha_{ji}(t) \neq 0$ satisfy

$$C_j^\dagger \alpha_{ji}(t) \in \ker[A_{ij}] \quad (3.22)$$

for all $t \geq T_a$ for it to be structurally stealthy to reduced-information detector $\mathcal{D}_i$. $\square$

Corollary 3.2 highlights how it is possible, in certain cases, for malicious agents to design structurally stealthy attacks for reduced-information detectors by relying on knowledge of the structure of the interconnection between subsystems. This is not only more easily obtained, as it depends on the physical relationships linking subsystems' dynamics, but it is particularly pernicious for LSSs where, as pointed out in [Lun92], subsystems are often weakly coupled, and therefore their interconnections may be structural-rank deficient.

This is particularly important for the case study evaluated here, i.e. that of DC microgrids. Indeed, recalling the definition of the interconnection matrices A_{ij} in Chapter 2.1.4, as well as the secondary control layer governing the consensus-based current sharing protocol, any attack on the terminal current I_{tji}^c , i.e. the measurement of the terminal current I_{tj} as received by $\mathcal{C}_i$, satisfies condition of Corollary 3.2. It is therefore paramount to include a method to systematically address the problem of structurally-stealthy attacks.

Given the possibility of structurally stealthy attacks, different approaches may be taken to make the control insensitive to these attacks. For example, by designing the distributed control law (3.6) such that

$$\kappa_i(0, \bar{\gamma}_{ji}^c) = 0 \quad (3.23)$$

is satisfied, where the decomposition $\gamma_{ji}^c = \bar{\gamma}_{ji}^c + \gamma_{ji}^{c\perp}$ is such that $\bar{\gamma}_{ji}^c \in \ker A_{ij}$, for all $j \in \mathcal{N}_i$. Indeed, if such a controller can be designed, any structurally stealthy attack to $\mathcal{D}_i$ is made

harmless, as it has no effect on the system.

If this is not possible, as in the case of DC microgrids, another approach must be developed. One such possibility is to expand the information set $\mathcal{I}_i$, allowing for the inclusion of information on the dynamics of $\mathcal{S}_j, j \in \mathcal{N}_i$. By doing so, recalling Remark 3.7, we can alter the class of attacks that are structurally stealthy to $\mathcal{D}_i$, possibly eliminating them.

In the remainder of this chapter, we focus on the design and the properties of a distributed diagnoser $\mathcal{D}_i$ which is not reduced-information, but rather is designed using knowledge of the dynamics of its neighbouring subsystems $\mathcal{S}_j, j \in \mathcal{N}_i$.

3.3 Local diagnoser $\mathcal{D}_i$ – A multi-module design

In this section, we evaluate the overall architecture of the distributed diagnostic module $\mathcal{D}_i$, without focusing on its properties, which are the subject of later sections. The distributed monitoring module $\mathcal{D}_i$, shown in Figure 3.2, is composed of two parallel modules:

- a distributed Luenberger-like observer $\mathcal{O}_i^L$, estimating the local state x_i designed with the reduced information set in Definition 3.5;
- a bank of N_i unknown-input observers (UIO) $\mathcal{O}_{ji}^{UIO}$, estimating the states x_j of neighbouring subsystems $\mathcal{S}_j, j \in \mathcal{N}_i$, designed with knowledge of their dynamical model.

As shown in the following, these modules are capable of exploiting different, and complementary, *trusted information* to detect cyber-attacks on $\gamma_{ji,j \in \mathcal{N}_i}^c$.

Definition 3.8 (Trusted information). *Any information available to $\mathcal{D}_i$ either during online or offline operations is considered trusted if it cannot be altered by a malicious agent.* $\triangleleft$

Remark 3.8. Note that the *trusted information* of any diagnoser $\mathcal{D}$ is necessarily a subset of its *information set* [Definition 3.4]. $\triangleleft$

As a matter of fact, module $\mathcal{O}_i^L$ relies on local trusted information and the dynamic coupling between x_i and $x_j, j \in \mathcal{N}_i$, to perform detection. Indeed, given that local measurements y_i and

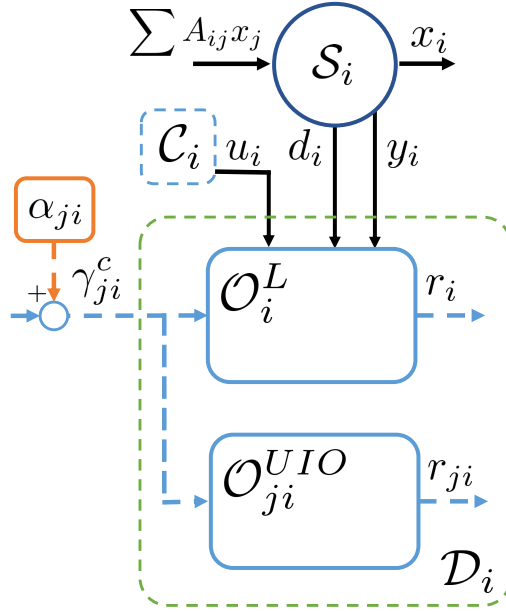


Figure 3.2: Representation of the diagnoser $\mathcal{D}_i$ used to detect cyber attacks. The detection module $\mathcal{D}_i$ receives measurements y_i , and exogenous inputs d_i from subsystem $\mathcal{S}_i$, as well as control inputs u_i from controllers $\mathcal{C}_i$. The received information γ_{ji}^c is subject to additive attack signal α_{ji} . The two modules $\mathcal{O}_i^L$ and $\mathcal{O}_{ji}^{UIO}$ respectively receive $y_i, d_i, u_i, \gamma_{ji, j \in \mathcal{N}_i}$, and $\gamma_{ji, j \in \mathcal{N}_i}$ and generate residuals r_i and r_{ji} .

inputs u_i are trusted by assumption (recall the attack model (3.7) and Assumption 3.7), the effect of ξ_i on the local dynamics can be compared with an estimate $\hat{\xi}_i$, computed with received information $\gamma_{ji}^c, j \in \mathcal{N}_i$. Thus, module $\mathcal{O}_i^L$ exploits the physical coupling between subsystems to detect attacks. Given this design choice, $\mathcal{O}_i^L$ can be seen as a reduced information module, and therefore there are attacks that are structurally stealthy to it, as seen in Section 3.2.

Conversely, by relying only on γ_{ji}^c (indeed, as is shown in Section 3.5, the dynamics of $\hat{x}_{ji}$ do not depend on y_i or u_i), the module $\mathcal{O}_{ji}^{UIO}$ compares the received information γ_{ji}^c with its expected trajectory, given knowledge of the model of $\mathcal{S}_j$. As such, $\mathcal{O}_{ji}^{UIO}$ presents vulnerabilities similar to those of centralised detection strategies⁶, as summarised in [TSSJ15] and Chapter 2. In fact, $\mathcal{S}_j$ – isolated from the rest of the LSS – can be seen as a stand-alone plant, with $\mathcal{O}_{ji}^{UIO}$ its centralized diagnosis module.

Although both modules have shortcomings, these are shown to be complementary to each other. Thus, by combining the two together and having them run in parallel and simultaneously, the diagnostic module $\mathcal{D}_i$ can detect a wider set of attack strategies than either of its modules on their own. Indeed, as formally shown in Section 3.6, for any attack $\alpha_{ji} \neq 0$ to be stealthy

⁶This is addressed in detail in Section 3.5.

to $\mathcal{D}_i$ it must be stealthy to *both* $\mathcal{O}_i^L$ and $\mathcal{O}_{ji}^{UIO}$, as if either (3.28a) or (3.28b) hold, the null hypothesis $\mathcal{H}_i^0$ in Problem 3.1 is violated and thus the attack is detected.

Let us give a brief overview of the main characteristics of both $\mathcal{O}_i^L$ and $\mathcal{O}_{ji}^{UIO}$, the details of which are presented in Sections 3.4 and 3.5, respectively.

3.3.1 $\mathcal{O}_i^L$ – An overview

The module $\mathcal{O}_i^L$ generates an estimate $\hat{x}_i$ capable of tracking the state of $\mathcal{S}_i$ in nominal conditions. Specifically, the estimation error $e_i \doteq x_i - \hat{x}_i$, in absence of attacks on the communication network and of any disturbances on (3.1), satisfies

$$\lim_{t \rightarrow \infty} |e_i(t)| = \lim_{t \rightarrow \infty} |x_i(t) - \hat{x}_i(t)| = 0, \quad (3.24)$$

exploiting – for example and without loss of generality – a Luenberger-like distributed observer.

In order to account for the physical coupling between $\mathcal{S}_i$ and its neighbours, this observer relies on knowledge of the interconnection matrices⁷ A_{ij} , as well as on $\gamma_{ji, j \in \mathcal{N}_i}^c$ received from its neighbours. The estimated state is then used to generate a residual, defined as

$$r_i(t) \doteq y_i(t) - C_i \hat{x}_i(t) \quad (3.25)$$

compared with an appropriately defined threshold $\bar{r}_i$. Indeed, given the design of the estimator, the deterministic nature of the dynamics, and the assumption of boundedness of the system and measurement disturbances in Assumption 3.1, it is possible to define a time-varying bound $\bar{r}_i$ which guarantees that the inequality

$$|r_i(t)| \leq \bar{r}_i(t) \quad (3.26)$$

holds component-by-component for all $t \geq 0$ and for any realization of the disturbances w_i and v_i satisfying the boundedness assumption, so long as $\alpha_{ji} = 0, j \in \mathcal{N}_i$.

⁷Note that perfect knowledge of A_{ij} may not be available. However, we consider that any uncertainty on the dynamics can be included in w_i .

Remark 3.9. Note that, in Section 3.2, we had assumed $\bar{w}_i = 0$ and $\bar{v}_i = 0$, as we wished to analyse the properties of a diagnoser $\mathcal{D}_i$ with respect to the information set $\mathcal{I}_i$ used for its design. On the other hand, here we consider $\bar{w}_i \neq 0$ and $\bar{v}_i \neq 0$, as we are also interested in analysing the performance of $\mathcal{D}_i$ with respect to the bounds on process and measurement disturbances. $\triangleleft$

Given its design, the module is capable of detecting attacks on the communication network whenever their effect on the estimated state dynamics is “sufficiently” different from the effect the neighbours’ states have on x_i , as highlighted in Section 3.4.

3.3.2 $\mathcal{O}_{ji}^{UIO}$ – An overview

Differently from $\mathcal{O}_i^L$, $\mathcal{O}_{ji}^{UIO}$ is designed to generate an estimate $\hat{x}_{ji}$ of state x_j , $j \in \mathcal{N}_i$, such that the estimation error $e_{ji}(t) \doteq x_j(t) - \hat{x}_{ji}(t)$ asymptotically converges to zero as $t \rightarrow \infty$, in absence of attacks and without disturbances.

Specifically, we equip $\mathcal{O}_{ji}^{UIO}$ with unknown-input observers [CPZ96] to estimate x_j , thus limiting the amount of information required to be received by $\mathcal{C}_i$ and $\mathcal{D}_i$ from $\mathcal{S}_j, j \in \mathcal{N}_i$. In fact, by design, this class of observers is capable of asymptotically estimating the state of an LTI system whose dynamics are influenced by unknown inputs, so long as certain constraints are satisfied. A summary of the design and operation of UIOs in a centralised framework is presented in Appendix A.

Similarly to $\mathcal{O}_i^L$, given asymptotic convergence of the estimation error, together with Assumption 3.1, it is possible to define a time-varying threshold $\bar{r}_{ji}$ such that, in nominal conditions, for residual $r_{ji} \doteq \gamma_{ji}^c - C_j \hat{x}_{ji}$, the condition

$$|r_{ji}(t)| \leq \bar{r}_{ji}(t) \quad (3.27)$$

holds for all $t \geq 0$ and all realizations of $w_j(t)$ and $v_j(t)$ satisfying Assumption 3.1.

Remark 3.10 (On the conservativeness of detection thresholds). The bounds $\bar{r}_i$ and $\bar{r}_{ji}$ in (3.26) and (3.27) are designed to be “conservative”, i.e. to hold for all t and all realisations

of w_i and v_i satisfying Assumption 3.1. While not detecting all attacks, this design strategy guarantees that operations of the plant are not altered due to false alarms. This is particularly important in systems where the interruption of operations without verified cause may lead to high costs, whereas the operation under malfunction may be less costly. $\triangleleft$

Having briefly introduced the *ingredients* that make up $\mathcal{D}_i$, we can summarise the design and operation of the detection monitor itself, and provide an overview of its properties.

3.3.3 $\mathcal{D}_i$ – Online operations and offline design

For the sake of clarity, we separate the description of the information required offline for the design of the modules within $\mathcal{D}_i$ from the online operation of $\mathcal{D}_i$, as shown in Table 3.2.

The online operation of each $\mathcal{D}_i$, $i \in \mathcal{N}$ is based on continuously updating estimates $\hat{x}_i$ and $\hat{x}_{ji}$, $j \in \mathcal{N}_i$, as well as the thresholds $\bar{r}_i$ and $\bar{r}_{ji}$, $j \in \mathcal{N}_i$. Hence, the detection tests

$$|r_i(t)| > \bar{r}_i(t) \tag{3.28a}$$

$$|r_{ji}(t)| > \bar{r}_{ji}(t) \tag{3.28b}$$

can be evaluated. If either (3.28a) or (3.28b) hold, the null hypothesis $\mathcal{H}_i^0$ in Problem 3.1 is rejected, thus detecting an attack $\alpha_{ji} \neq 0$. Moreover, by nature of the design of the two modules, if (3.28b) is violated at any time⁸ $t > \tilde{T}_a$, $\tilde{T}_a \doteq \min_{j \in \mathcal{N}_i} T_a^{ji}$, the communication line over which the attack is active is isolated, as further discussed in Section 3.7, where the problem of *attack isolation* is formally introduced. The overall nominal operation of $\mathcal{D}_i$ is summarised in Algorithm 1.

Further to the description of how $\mathcal{D}_i$ operates in nominal operations, and how it is used to detect attacks, it is important to describe what information is required by $\mathcal{O}_i^L$ and $\mathcal{O}_{ji}^{UIO}$ for their design, and what is required for their online operations.

For offline design, both estimators require knowledge of the dynamics of the subsystem whose state is being estimated, namely $\mathcal{S}_i$ for $\mathcal{O}_i^L$ and $\mathcal{S}_j$ for $\mathcal{O}_{ji}^{UIO}$, $j \in \mathcal{N}_i$ from (3.1). For

⁸For the remainder of this chapter we consider that only $\mathcal{S}_i$ is under attack, and therefore omit the index i related to $\tilde{T}_a$ for legibility.

Algorithm 1 Attack detection at time t

```

1: while  $\mathcal{S}_i$  online  $\forall i \in \mathcal{N}$  do
2:   Receive  $\gamma_{ji}^c, j \in \mathcal{N}_i$  from neighbours  $\mathcal{S}_j, j \in \mathcal{N}_i$ ;
3:   Update estimate  $\hat{x}_{ji}(t), j \in \mathcal{N}_i$ , given  $\gamma_{ji}^c$ ;
4:   Update estimate  $\hat{x}_i(t)$  given  $y_i, u_i, \gamma_{ji}^c, j \in \mathcal{N}_i$ ;
5:   Compute residuals  $r_i(t)$  and  $r_{ji}(t), j \in \mathcal{N}_i$ ;
6:   Update bounds  $\bar{r}_i(t)$  and  $\bar{r}_{ji}(t), \forall j \in \mathcal{N}_i$ ;
7:   Evaluate (3.28a) and (3.28b)
8:   if (3.28a) or (3.28b) hold then
9:     if  $|r_{ji}(t)| > \bar{r}_{ji}(t)$  for any  $j \in \mathcal{N}_i$  then
10:      Attack detected on link  $(j, i)$ 
11:     else
12:       Attack detected
13:     end if
14:   else
15:     No attack is detected at time  $t$ 
16:   end if
17: end while

```

Table 3.2: Information required for design of $\mathcal{D}_i$ and attack detection

	$\mathcal{O}_{ji}^{UIO}, \forall j \in \mathcal{N}_i$	$\mathcal{O}_i^L$
Offline Information	Matrices $\mathbf{A}_{jj}, \mathbf{E}_j, \mathbf{C}_j$, bounds $\bar{w}_{[j]}$ and $\bar{v}_{[j]}$	Model of $\mathcal{S}_i$, matrix $\mathbf{C}_j$, and bounds $\bar{w}_{[i]}$ and $\bar{v}_{[k]}$, for all $k \in \{i\} \cup \mathcal{N}_i$
Online Information	Communicated measurements $\gamma_{ji}^c(t)$	Local measurements and inputs, and communicated measurements $\gamma_{ji}^c(t)$

online operations $\mathcal{O}_i^L$ requires y_i, u_i and received information $\gamma_{ji}^c, j \in \mathcal{N}_i$ to compute update of $\hat{x}_i$, while $\mathcal{O}_{ji}^{UIO}$ needs γ_{ji}^c , to compute $\hat{x}_{ji}$. Furthermore, to compute detection threshold $\bar{r}_i$, $\mathcal{D}_i$ requires uncertainty bounds $\bar{w}_i$ and $\bar{v}_i$, while for $\bar{r}_{ji}, \bar{w}_j, \bar{v}_j, j \in \mathcal{N}_i$ are required. The thresholds $\bar{r}_i$ and $\bar{r}_{ji}$ are defined in Section 3.4 and Section 3.5, respectively, to guarantee cyber-attack detection without any false alarms. Table 3.2 summarises the information needed.

3.3.4 A note on the topology of the communication network

Throughout the remainder of this chapter, we consider that the communication network introduced for control, and the network given by the physical coupling between subsystems have the same topology, as expressed in Assumption 3.4. This may not always be the case. Hence, let us address how the diagnoser $\mathcal{D}_i$ may be designed for LSSs that do not satisfy Assumption 3.4.

Recall from Assumption 3.4 that the two networks can be represented as two graphs, defined as $\mathcal{G}^p \doteq \{\mathcal{V}^p, \mathcal{E}^p\}$ and $\mathcal{G}^c \doteq \{\mathcal{V}^c, \mathcal{E}^c\}$, respectively for the physical and communication graphs, where $\mathcal{V}^s$ is the set of nodes and $\mathcal{E}^s$ is the set of edges, for $s = \{p, c\}$. The set of nodes of the graphs are identical, $\mathcal{V}^p = \mathcal{V}^c = \mathcal{N}$, while the set of edges may differ. Let us assume the following:

Assumption 3.9. *The communication network is such that $\mathcal{E}^p \subseteq \mathcal{E}^c$ holds.* $\triangleleft$

Under Assumption 3.9, the set of neighbours induced by the communication graph $\mathcal{N}_i^c$ is such that $\mathcal{N}_i^p \subseteq \mathcal{N}_i^c$, where $\mathcal{N}_i^p \equiv \mathcal{N}_i$. As such, to design the diagnoser to monitor all information γ_{ji}^c received from neighbours in $\mathcal{N}_i^c$, while maintaining the construction of $\mathcal{O}_i^L$ without alteration, additional $\mathcal{O}_{ji}^{UIO}$ must be included for all $j \in \mathcal{N}_i^c \setminus \mathcal{N}_i^p$.

The diagnoser $\mathcal{D}_i$ thus maintains its detection properties for all those neighbours which are physically coupled with $\mathcal{S}_i$. However, when detecting attacks on received information $\gamma_{ji}^c, j \in \mathcal{N}_i^c \setminus \mathcal{N}_i^p$, the module $\mathcal{O}_i^L$ does not provide any additional support to the diagnosis. Therefore, for detection on this set of information, the properties of $\mathcal{D}_i$ are equivalent to those of $\mathcal{O}_{ji}^{UIO}$, as presented in Section 3.5.

We now present the design and properties of $\mathcal{O}_i^L$ and $\mathcal{O}_{ji}^{UIO}$ independently, considering once again that Assumption 3.4 holds, i.e. assuming that $\mathcal{N}_i^p = \mathcal{N}_i^c = \mathcal{N}_i$.

3.4 Local state observer – $\mathcal{O}_i^L$

We now introduce the implementation of the local state estimation module $\mathcal{O}_i^L$. As previously mentioned, this module is based on the development of a distributed Luenberger-based observer to generate the estimate of the state $\hat{x}_i$.

3.4.1 Design of $\mathcal{O}_i^L$

For each $\mathcal{D}_i$, the state estimate $\hat{x}_i$ has dynamics:

$$\hat{\mathcal{S}}_i : \begin{cases} \dot{\hat{x}}_i = A_{ii}\hat{x}_i + B_i u_i + \hat{\xi}_i + E_i d_i - L_i(y_i - \hat{y}_i) \\ \hat{y}_i = C_i \hat{x}_i \end{cases} \quad (3.29)$$

where $\hat{x}_i \in \mathbb{R}^{n_i}$, $\hat{y}_i \in \mathbb{R}^{p_i}$ are the state and output estimates, respectively, $\hat{\xi}_i$ is the local estimate of the aggregate physical coupling with neighbours $\mathcal{N}_i$, and L_i is such that $A_{Li} \doteq A_{ii} + L_i C_i$ is Hurwitz stable. Note that existence of a stabilizing matrix L_i follows from Assumption 3.2, guaranteeing stability of the estimation error. Estimate $\hat{\xi}_i$ is defined as:

$$\hat{\xi}_i \doteq \sum_{j \in \mathcal{N}_i} A_{ij} C_j^\dagger \gamma_{ji}^c \quad (3.30)$$

with $C_j^\dagger \gamma_{ji}^c$ used as an estimate of x_j .

Remark 3.11. Note that, given Assumption 3.3, $\text{rank } C_j = p_j < n_j$ does not introduce any error on the local estimate of the dynamic physical interconnection between $\mathcal{S}_i$ and its neighbours, as $x_j \in \ker(C_j) \implies x_j \in \ker(A_{ij})$. Hence,

$$\hat{\xi}_i = \sum_{j \in \mathcal{N}_i} A_{ij} \left(x_j + C_j^\dagger v_j \right)$$

is equivalent to the definition of $\hat{\xi}_i$ in (3.30). Thus, $\hat{\xi}_i$ does not introduce any bias to the estimation error or residual. $\triangleleft$

As previously introduced, to detect attacks, the residual

$$r_i \doteq y_i - \hat{y}_i \quad (3.31)$$

is computed and compared with the detection threshold $\bar{r}_i$, defined as

$$\bar{r}_i \doteq C_i e^{-\lambda_i t} \Lambda_i \bar{\epsilon}_i^0 + C_i \int_0^t e^{-\lambda_i(t-\tau)} \Lambda_i \bar{\eta}_i d\tau + \bar{v}_i \quad (3.32)$$

where $\bar{\epsilon}_i^0 \geq 0$ is an appropriately defined bound on the unknown initial error $|\epsilon_i(0)| = |x_i(0) - \hat{x}_i(0)|$; moreover, $\bar{\eta}_i \doteq \bar{w}_i + |L_i| \bar{v}_i + \sum_{j \in \mathcal{N}_i} |A_{ij}| C_j^\dagger \bar{v}_j$; finally, $\lambda_i > 0$ and $\Lambda_i \geq I$ are such that $|e^{A_{Li}t}| \leq e^{-\lambda_i t} \Lambda_i, \forall t \geq 0$, as per the following lemma. Note that $\bar{r}_i$ is a vector threshold that is compared component by component to $|r_i|$ in the detection test (3.28a). As such, as is made apparent in the following, it is necessary to find scalar $\lambda_i > 0$ and matrix $\Lambda_i \geq I$ such that $|e^{A_{Li}t}| \leq e^{-\lambda_i t} \Lambda_i, \forall t \geq 0$ holds elementwise. Recall that with the operator $|\cdot|$ applied to vectors

or matrices, we intend the element-by-element absolute value. Let us note that, although not a norm, the element-by-element absolute value operator satisfies the following properties for any two matrices A and B of appropriate dimensions:

- $|A + B| \leq |A| + |B|$;
- $|AB| \leq |A||B|$.

For all $i \in \mathcal{N}$, $C_i \geq 0$ is assumed without loss of generality.

Lemma 3.1. If a matrix $A \in \mathbb{R}^{n \times n}$ is Hurwitz stable, there exist a positive scalar $\kappa > 0$ and a matrix $K \in \mathbb{R}^{n \times n}$, $K \geq I$, such that:

$$|e^{At}| \leq e^{-\kappa t} K \quad (3.33)$$

holds for all $t \geq 0$. □

Proof. For any matrix A , we define its Jordan normal form $J \doteq P^{-1}AP$. This implies the equivalence

$$e^{At} = Pe^{Jt}P^{-1}.$$

Note that e^{Jt} has the same block-diagonal structure of J , and therefore each block $e^{J_k t} \in \mathbb{R}^{n_k \times n_k}$ is upper-triangular. Following this, define a block-diagonal matrix $S \doteq \text{diag}_k(S_k)$, such that each block $S_k \in \mathbb{R}^{n_k \times n_k}$ of the form

$$S_k \doteq \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 0 & 1 & \cdots & 1 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \end{bmatrix} \quad (3.34)$$

has the same non-zero structure as $e^{J_k t}$. Therefore, the matrix S has the same non-zero structure

as e^{Jt} and it may be rewritten as $S = I + \Xi$, with $\Xi = \text{diag}(\Xi_k)$, and $\Xi_k \in \mathbb{R}^{n_k \times n_k}$ defined as:

$$\Xi_k \doteq \begin{bmatrix} 0 & 1 & \cdots & 1 \\ 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & 0 \end{bmatrix}. \quad (3.35)$$

Given the structure of J and that of S defined above, as well as the property of the Euclidian norm that for any matrix $M = m_{ij}$, $M \in \mathbb{R}^{n \times n}$

$$\max_{i,j \in \{1, \dots, n\}} |m_{ij}| \leq \|M\| \quad (3.36)$$

and thus the inequality

$$|Pe^{Jt}P^{-1}| \leq |P| |e^{Jt}| |P^{-1}| \leq \|e^{Jt}\| |P|S|P^{-1}| \quad (3.37)$$

is satisfied elementwise, given that

$$|e^{Jt}| \leq \|e^{Jt}\| S. \quad (3.38)$$

Hence, noting that if A is Hurwitz stable then J is also Hurwitz stable, it is possible to find scalars $\kappa > 0$ and $\mu \geq 1$ [HL04] such that:

$$\|e^{Jt}\| \leq \mu e^{-\kappa t}. \quad (3.39)$$

and to define $K \doteq \mu |P|S|P^{-1}|$.

Finally, note that matrix K is such that $K \geq I$ holds, as the following relationships hold:

$$\begin{aligned} |P|S|P^{-1}| &= |P|(I + \Xi)|P^{-1}| \\ &= |P||P^{-1}| + |P|\Xi|P^{-1}| \geq I + 0, \end{aligned} \quad (3.40)$$

given that $\Xi \geq 0$ by definition, and that $|P||P^{-1}| \geq |PP^{-1}| = I$, $|P|\Xi|P^{-1}| \geq 0$. ■

Proposition 3.1 (False alarms in $\mathcal{O}_i^L$). Given Assumption 3.1, Hurwitz stability of A_{Li} , and

the definition of the residual threshold $\bar{r}_i(t)$ in (3.32), the inequality

$$|r_i(t)| \leq \bar{r}_i(t) \quad (3.41)$$

holds for all $t \leq T_a^{ji}, j \in \mathcal{N}_i$, independently of $\epsilon_i(0)$, and all $w_i(s), v_i(s), v_j(s), \forall s \in [0, t]$. $\square$

Proof. First, we note that the residual can be rewritten as $r_i = C_i \epsilon_i + v_i$, with $\epsilon_i \doteq x_i - \hat{x}_i$. Furthermore, the dynamics of the estimation error can be found to be:

$$\dot{\epsilon}_i = A_{Li} \epsilon_i - \sum_{j \in \mathcal{N}_i} A_{ij} C_j^\dagger v_j + w_i - L_i v_i. \quad (3.42)$$

The explicit solution of (3.42) is given by:

$$\epsilon_i(t) = e^{A_{Li}t} \epsilon_i(0) + \int_0^t e^{A_{Li}(t-\tau)} \left[- \sum_{j \in \mathcal{N}_i} A_{ij} C_j^\dagger v_j(\tau) + w_i(\tau) - L_i v_i(\tau) \right] d\tau. \quad (3.43)$$

Given the Hurwitz stability of A_{Li} , by construction, from Lemma 3.1 a matrix $\Lambda_i \geq I$ and a scalar $\lambda_i \geq 0$ can be found such that $|e^{A_{Li}t}| \leq \Lambda_i e^{-\lambda_i t}, \forall t \geq 0$ holds, component-by-component. Applying the triangle inequality, and recalling Assumption 3.1 and the construction of γ_{ji}^c , in nominal operations $\epsilon_i(t)$ is bounded for all t by $\bar{\epsilon}_i$, defined by:

$$\bar{\epsilon}_i = e^{-\lambda_i t} \Lambda_i \bar{\epsilon}_i^0 + \int_0^t e^{-\lambda_i(t-\tau)} \Lambda_i \left[\sum_{j \in \mathcal{N}_i} |A_{ij}| C_j^\dagger \bar{v}_j(\tau) + \bar{w}_i(\tau) + |L_i| \bar{v}_i(\tau) \right] d\tau. \quad (3.44)$$

where $\bar{\epsilon}_i^0 \geq 0 \in \mathbb{R}^{n_i}$ is appropriately defined to bound $|\epsilon_i(0)|$. Hence, again by use of the triangle inequality, bound (3.32) is computed, thus showing it is an upper bound of r_i for all $t \geq 0$, irrespective of the magnitude of w_i, v_i, v_j satisfying Assumption 3.1. $\blacksquare$

Remark 3.12. Noting that the residual threshold $\bar{r}_i$ is computed by using the worst case bounds on the uncertainties w_i, v_i , and v_j , it guarantees that whenever inequality (3.28a) does not hold, there is an attack active on the communication between $\mathcal{S}_i$ and one of its neighbours. Although this allows to account for any unstructured disturbance affecting either the dynamics or the measurements, it is *by definition* a worst case scenario, and as such conservative.

If any further information is known about the disturbance signals (or if they can be described as stochastic signals, as shown in Chapter 4), this may be included in the design of the residual bound to make it less conservative, at the price of introducing a nonzero probability of false alarm. $\triangleleft$

3.4.2 Properties of local state observer $\mathcal{O}_i^L$

Having presented the structure of $\mathcal{O}_i^L$, we now focus on its properties while under attack, i.e. for $t \geq \check{T}_a$. We analyze both *detectable* and *stealthy* attacks defined in Chapter 2. Once an attack is initialised, it affects the local Luenberger-based module $\mathcal{O}_i^L$ through its estimate $\hat{\xi}_i$ of the physical coupling between $\mathcal{S}_i$ and its neighbours. Indeed⁹, for $t \geq \check{T}_a$:

$$\hat{\xi}_i(t) = \sum_{j \in \mathcal{N}_i} A_{ij} \left(x_j(t) + C_j^\dagger v_j(t) \right) + \sum_{j \in \hat{\mathcal{N}}_i(t)} A_{ij} C_j^\dagger \alpha_{ji}(t)$$

where $\hat{\mathcal{N}}_i(t) \doteq \{j \in \mathcal{N}_i : t \geq T_a^{ji}\} \subseteq \mathcal{N}_i$ is the set of neighbours for which the information transmitted to $\mathcal{S}_i$ is under attack at time t . Clearly, for $t < \check{T}_a$, $\hat{\mathcal{N}}_i = \emptyset$.

Given the linearity of the subsystem dynamics, the residual can be seen as the superposition of a “healthy” and an “attacked” component, namely $r_i = r_i^h + r_i^a$, where r_i^h is defined as

$$r_i^h = C_i e^{A_{Li}t} \epsilon_i(0) + \int_0^t e^{A_{Li}(t-\tau)} \left[- \sum_{j \in \mathcal{N}_i} A_{ij} C_j^\dagger v_j(\tau) + w_i(\tau) - L_i v_i(\tau) \right] d\tau + v_i, \quad (3.45)$$

and is independent of $\check{T}_a$. As such, given Proposition 3.1 and definition of r_i^h ,

$$|r_i^h(t)| \leq \bar{r}_i(t), \quad \forall t \geq 0 \quad (3.46)$$

holds by design of $\bar{r}_i$. On the other hand,

$$r_i^a(t) = \int_{\check{T}_a}^t e^{A_{Li}(t-\tau)} \left[- \sum_{j \in \hat{\mathcal{N}}_i} A_{ij} C_j^\dagger \alpha_{ji}(\tau) \right] d\tau \quad (3.47)$$

⁹Note that, here and throughout the remainder of this Chapter, the term $\beta_{ji}(t - T_a^{ji})$ is dropped for the sake of simplifying the notation. Given that $\beta_{ji}(t) \in [0, 1], \forall t \geq 0$, and $\beta_{ji}(t) \rightarrow 1$ monotonically as $t \rightarrow \infty$, this approximation does not influence the generality of the results.

is the component of $r_i(t)$ under attack. Note that any alteration of u_i caused by the attack, although affecting the trajectory of x_i , does not affect r_i^a , as it influences both x_i and $\hat{x}_i$ in the same way.

Recalling the definitions of detectable and stealthy attacks in Chapter 2, we now analyze the main detectability properties of $\mathcal{O}_i^L$.

Proposition 3.2 (Detectable attacks). For a subsystem $\mathcal{S}_i$ at time $t \geq \check{T}_a$, if attacks $\alpha_{ji}(s)$, $j \in \hat{\mathcal{N}}_i(s)$, $s \in [\check{T}_a, T_d^i]$, are such that the inequality

$$|r_i^a(T_d^i)| \geq 2\bar{r}_i(T_d^i) \quad (3.48)$$

is satisfied for any component, then the attack is detected by $\mathcal{O}_i^L$ at some time $t \leq T_d^i$. $\square$

Proof. Using the inverse triangle inequality and (3.46) one sees that

$$|r_i| = |r_i^h + r_i^a| \geq |r_i^a| - |r_i^h| \geq |r_i^a| - \bar{r}_i \quad (3.49)$$

holds. Thus, (3.48) is a sufficient condition for an attack to be detected by $\mathcal{O}_i^L$. $\blacksquare$

Remark 3.13. Note that Proposition 3.2 represents a worst case guarantee of detection, being a sufficient condition for detection. $\triangleleft$

Proposition 3.3 (Stealthy attacks). If the attack functions $\alpha_{ji}(t)$, $j \in \hat{\mathcal{N}}_i$ satisfy

$$\alpha_i(t) \doteq \text{col}_{j \in \hat{\mathcal{N}}_i} \left(C_j^\top \alpha_{ji}(t) \right) \in \ker \text{row}_{j \in \hat{\mathcal{N}}_i} (A_{ij}) \quad \forall t \geq \check{T}_a \quad (3.50)$$

then the attack is stealthy. $\square$

Proof. Given that $|r_i^h(t)| \leq \bar{r}_i(t)$, $\forall t \geq 0$ by construction of $\bar{r}_i(t)$, one gets

$$|r_i| = |r_i^h + r_i^a| \leq |r_i^a| + \bar{r}_i. \quad (3.51)$$

As such, given that for detection to be avoided in $\mathcal{O}_i^L$ for all t (3.28a) must always hold, (3.50) is a sufficient condition for an attack to be undetected, by definition of r_i^a in (3.47). $\blacksquare$

As seen from Proposition 3.2 and Proposition 3.3, the properties of $\mathcal{O}_i^L$ are tightly linked to the physical coupling between $\mathcal{S}_i$ and its neighbours, also reflecting the main result in Section 3.2, given that $\mathcal{O}_i^L$ can be seen as a reduced-information distributed monitoring module.

Furthermore, Proposition 3.3 demonstrates the vulnerability of $\mathcal{O}_i^L$ towards coordinated cyber-attacks on multiple communication lines. Indeed, even if x_j were input observable from y_i , there exists some minimum size of $\widehat{\mathcal{N}}_i(t)$ above which it is possible for the malicious agent to construct a signal satisfying (3.50), modulo knowledge of the interconnection matrices A_{ij} .

3.4.3 Using state estimates from $\mathcal{O}_{ji}^{UIO}$ to compute $\hat{x}_i$

From the analysis of the properties of $\mathcal{O}_i^L$ presented thus far, it emerges that the detection module relies on the physical coupling between subsystems. As such, it is fundamental that information on the dynamic coupling is contained in the received signal γ_{ji}^c , as shown by Assumption 3.3.

If C_j is not such that Assumption 3.3 holds, while still satisfying Assumption 3.10, the unknown-input observers in $\mathcal{O}_{ji}^{UIO}$ may still be constructed. Hence, the dynamics of the local state observer (3.29) can be modified by substituting γ_{ji}^c with $\hat{x}_{ji}$ in the estimate of physical coupling:

$$\hat{\xi}_i = \sum_{j \in \mathcal{N}_i} A_{ij} \hat{x}_{ji} = \sum_{j \in \mathcal{N}_i} A_{ij} (x_j + \epsilon_{ji}), \quad (3.52)$$

where $\hat{x}_{ji}$ is the state estimate of x_j as computed by the module $\mathcal{O}_{ji}^{UIO}$, and $\epsilon_{ji} \doteq x_j - \hat{x}_{ji}$ is its estimation error.

Following this substitution, the properties of $\mathcal{O}_i^L$ expressed in Proposition 3.2 and Proposition 3.3 remain unchanged, once $\bar{\eta}_i$ in (3.32) is redefined as:

$$\bar{\eta}_i = \bar{w}_i + |L_i| \bar{v}_i + \sum_{j \in \mathcal{N}_i} |A_{ij}| \bar{\epsilon}_{ji}, \quad (3.53)$$

where $\bar{\epsilon}_{ji}, j \in \mathcal{N}_i$ is a bound on the estimation error ϵ_{ji} satisfying $|\epsilon_{ji}(t)| \leq \bar{\epsilon}_{ji}(t)$, for all t , in nominal conditions. Both $\hat{x}_{ji}$ and $\bar{\epsilon}_{ji}$ are formally introduced in the following.

3.5 Neighbour state observer – $\mathcal{O}_{ji}^{UIO}$

Having presented the structure and properties of $\mathcal{O}_i^L$, let us now introduce $\mathcal{O}_{ji}^{UIO}$. In terms of the information set presented in Chapter 3.2, by introducing this module we increase the model knowledge available to $\mathcal{D}_i$ to include the system dynamics of $\mathcal{S}_j$, for all the neighbours $j \in \mathcal{N}_i$ of $\mathcal{S}_i$. This requires the presumption that the information obtained *offline* for design is trustworthy, i.e. that the model information that each $\mathcal{S}_i$ has of its neighbours is correct. As anticipated previously, the main insight in the design of $\mathcal{O}_{ji}^{UIO}$ is the desire of monitoring the received signal γ_{ji}^c and thus checking its time-varying trajectory against its nominal model, i.e. the dynamics of $\mathcal{S}_j$.

As seen from (3.1), the dynamics of x_j are affected by a number of inputs, such as the subsystem's control input, the exogenous inputs acting upon $\mathcal{S}_j$, and the physical coupling with neighbours in $\mathcal{N}_j$. This implies that the use of a distributed Luenberger-like observer, such as $\mathcal{O}_i^L$, would require additional information to be transmitted between subsystems in real time. This is problematic, as any additional signal transmitted would also be exposed to attacks, and thus would need to be monitored, requiring more resources and information. To circumvent this, each $\mathcal{D}_i$ is equipped with a bank of N_i unknown-input observers (UIOs) to compute estimates of x_j .

As presented in Appendix A, UIOs are a class of observers designed to algebraically decouple the output estimation error from unknown inputs affecting the state dynamics [CPZ96]. As such they allow for the estimation of x_j by $\mathcal{D}_i$ without including further signals in the communication.

In the remainder of this section, we present the design of $\mathcal{O}_{ji}^{UIO}$, as well as its properties, described in terms of those classes of attacks which are guaranteed to be detected and those which are stealthy to the module. We show how, given that its design and operation are dependent exclusively on the dynamics of $\mathcal{S}_j$ and the signal γ_{ji}^c , $j \in \mathcal{N}_i$, the properties of $\mathcal{O}_{ji}^{UIO}$ have similar characteristics to those of a centralised cyber-attack diagnosis scheme, as given in Chapter 2.

3.5.1 Module design

To design $\mathcal{O}_{ji}^{UIO}$, it is necessary to express the dynamics of $\mathcal{S}_j$ isolating its dependence on all inputs unknown to $\mathcal{D}_i$. As such, let us rewrite (3.1) as:

$$\mathcal{S}_j : \begin{cases} \dot{x}_j = A_{jj}x_j + \bar{E}_j\bar{d}_j + w_j \\ y_j = C_jx_j + v_j \end{cases} \quad (3.54)$$

where the term $\bar{E}_j\bar{d}_j$ is such that

$$\bar{E}_j\bar{d}_j = B_ju_j + E_jd_j + \xi_j, \quad (3.55)$$

and therefore models the effect of the unknown inputs on the dynamics of x_j . The matrix $\bar{E}_j \in \mathbb{R}^{n_j \times q_j}$, with $q_j \leq n_j$, is designed such that its columns make up an independent basis of the column space

$$\text{Im} \begin{bmatrix} B_j & E_j & A_{jk_1} & \dots & A_{jk_{N_j}} \end{bmatrix}$$

where $\{k_1, \dots, k_{N_j}\} \equiv \mathcal{N}_j$. Thus, it is ensured that $\bar{E}_j$ is full-column rank, as required in [CPZ96] for UIO design. On the other hand, the vector $\bar{d}_j \in \mathbb{R}^{q_j}$ is defined as

$$\bar{d}_j \doteq \check{E}_j\check{d}_j,$$

i.e. it is a linear combination of $\check{d}_j$, the vector containing the inputs to $\mathcal{S}_j$ unknown to $\mathcal{D}_i$:

$$\check{d}_j \doteq \begin{bmatrix} u_j^\top & d_j^\top & x_{k_1}^\top & \dots & x_{k_{N_j}}^\top \end{bmatrix}^\top, \quad (3.56)$$

such that (3.55) holds. The full-order UIO is characterised by the dynamics:

$$\hat{\mathcal{S}}_{ji} : \begin{cases} \dot{z}_{ji} = F_jz_{ji} + \hat{K}_j\gamma_{ji}^c \\ \hat{x}_{ji} = z_{ji} + H_j\gamma_{ji}^c \\ \hat{\gamma}_{ji} = C_j\hat{x}_{ji} \end{cases} \quad (3.57)$$

where $\hat{z}_{ji} \in \mathbb{R}^{n_j}$ is the state of the observer, $\hat{x}_{ji} \in \mathbb{R}^{n_j}$ is the estimate of x_j computed by $\mathcal{O}_{ji}^{UIO}$, and $\hat{\gamma}_{ji}$ is the estimate of the received information γ_{ji}^c . The matrices in (3.57) satisfy the following:

$$(H_j C_j - I_n) \bar{E}_j = 0 \quad (3.58a)$$

$$S_j = I_n - H_j C_j \quad (3.58b)$$

$$F_j = S_j A_{jj} - \tilde{K}_j C_j \quad (3.58c)$$

$$\bar{K}_j = F_j H_j \quad (3.58d)$$

$$\hat{K}_j = \bar{K}_j + \tilde{K}_j \quad (3.58e)$$

as introduced in [CPZ96]. The dynamics (3.57) together with the design of matrices (3.58) is such that, without disturbances and in nominal conditions, the residual $r_{ji} \doteq \gamma_{ji}^c - \hat{\gamma}_{ji}$ satisfies

$$\lim_{t \rightarrow \infty} r_{ji} = 0$$

[CPZ96]. Further details on the properties of UIOs are presented in Appendix A.

The procedure needed to design UIO in $\mathcal{O}_{ji}^{UIO}$ can be summarised as follows: firstly, compute H_j satisfying (3.58a), and thus find S_j in (3.58b), allowing the estimate $\hat{x}_{ji}$ to be algebraically decoupled from the unknown input vector $\bar{d}_j$; secondly, define $\tilde{K}_j$ to guarantee stability of F_j in (3.58c), thus achieving asymptotic tracking independently of initial error and unknown inputs; finally, compute all other matrices in (3.58).

To satisfy the necessary conditions to design a UIO [Th. A.5], the following is assumed:

Assumption 3.10. *For all $i \in \mathcal{N}$, the conditions*

$$\text{rank}(C_i \bar{E}_i) = \text{rank}(\bar{E}_i), \quad (C3.1)$$

$$\text{the tuple } (A_{ii}, \bar{E}_i, C_i) \text{ is strongly observable} \quad (C3.2)$$

hold for the dynamics of $\mathcal{S}_i$ in (3.54).

◁

Because the design of the matrices (3.58) ensures stability of the estimation error $\epsilon_{ji} \doteq$

$x_j - \hat{x}_{ji}$, it is possible to design a bound on $|r_{ji}|$ to be used as a threshold in detection test (3.28b). Specifically, $\bar{r}_{ji}(t)$ is defined as:

$$\bar{r}_{ji}(t) = C_j e^{-\sigma_j t} \Sigma_j [\bar{\epsilon}_{ji}^0 + |H_j| \bar{v}_j] + |Z_j| \bar{v}_j + C_j \int_0^t e^{-\sigma_j(t-\tau)} \Sigma_j [|S_j| \bar{w}_j + |\hat{K}_j| \bar{v}_j] d\tau \quad (3.59)$$

where¹⁰ $Z_j = I_{p_j} - C_j H_j$; scalar $\sigma_j > 0$ and matrix $\Sigma_j \geq I$ can be found from Lemma 3.1, given Hurwitz stability of F_j ; $\bar{\epsilon}_{ji}^0 \geq 0$ is appropriately defined to bound initial estimation error $|\epsilon_{ji}(0)|$; bounds $\bar{w}_j$ and $\bar{v}_j$ are defined in Assumption 3.1. Proposition 3.4 demonstrates that $\bar{r}_{ji}$ is an appropriate upper bound for the residual r_{ji} , in the absence of an attack.

Proposition 3.4. Consider neighbouring subsystems $\mathcal{S}_i$ and $\mathcal{S}_j$, with dynamics (3.54) satisfying Assumption 3.10. Let $\mathcal{S}_i$ be equipped with diagnostic module $\mathcal{D}_i$ including $\mathcal{O}_{ji}^{UIO}$ to estimate the state x_j . Given the Hurwitz-stable matrix F_j designed as in (3.58), $\bar{r}_{ji}(t)$ as defined in (3.59) is such that the condition

$$|r_{ji}(t)| \leq \bar{r}_{ji}(t) \quad (3.60)$$

is guaranteed to hold for all $t \in [0, T_a^{ij})$, independently of $\epsilon_{ji}(0)$, and $w_j(s), v_j(s), \forall s \in [0, t]$. $\square$

Proof. Given dynamics of subsystem $\mathcal{S}_j$ in (3.54), dynamics of UIO $\mathcal{O}_{ji}^{UIO}$ (3.57), with matrices designed in (3.58), the solution of the estimation error $\epsilon_{ji}(t) \doteq x_j(t) - \hat{x}_{ji}(t)$ can be found to be [Appendix A]:

$$\epsilon_{ji}(t) = e^{F_j t} [\epsilon_{ji}(0) + H_j v_j(0)] - H_j v_j(t) + \int_0^t e^{F_j(t-\tau)} [S_j w_j(\tau) - \hat{K}_j v_j(\tau)] d\tau. \quad (3.61)$$

Given that $r_{ji}(t) = C_j \epsilon_{ji}(t) + v_j(t)$, the residual error at time $t < T_a^{ji}$ can be rewritten as:

$$r_{ji}(t) = C_j e^{F_j t} [\epsilon_{ji}(0) + H_j v_j(0)] + Z_j v_j(t) + C_j \int_0^t e^{F_j(t-\tau)} [S_j w_j(\tau) - \hat{K}_j v_j(\tau)] d\tau. \quad (3.62)$$

By use of Lemma 3.1 and Hurwitz stability of F_j , it is possible to find $\Sigma_j \geq I$, $\sigma_j \geq 0$ such that

$$|e^{F_j t}| \leq \Sigma_j e^{-\sigma t} \quad (3.63)$$

¹⁰Recall that $C_j \geq 0$ is supposed without loss of generality.

is satisfied elementwise. Thus, exploiting the bounds in Assumption 3.1 and the triangle inequality, we define a bound $\bar{\epsilon}_i(t)$ as

$$\bar{\epsilon}_{ji}(t) = e^{-\sigma_j t} \Sigma_j [\bar{\epsilon}_{ji}^0 + |H_j| \bar{v}_j(0)] + |H_j| \bar{v}_j + \int_0^t e^{-\sigma_j(t-\tau)} \Sigma_j \left[|S_j| \bar{w}_j(\tau) + |\hat{K}_j| \bar{v}_j(\tau) \right] d\tau, \quad (3.64)$$

satisfying

$$|\epsilon_{ji}(t)| \leq \bar{\epsilon}_{ji}$$

for all $t \geq 0$. Similarly, by further use of the triangle inequality, $\bar{r}_{ji}(t)$ defined in (3.59) is such that (3.60) holds for all $t \geq 0$. ■

Remark 3.14. It is worth noting that both $\bar{r}_i$ in (3.32) and $\bar{r}_{ji}$ in (3.59) are constructed in such a way that they can be seen as the output of a linear system. As such, they are easily implementable in real-world applications [ZPP04]. ◁

3.5.2 Detection properties

Similarly to Chapter 3.4, we now analyze the properties of $\mathcal{O}_{ji}^{UIO}$. Given that in this section we focus only on the signals received from one neighbouring subsystem $\mathcal{S}_j$, let us define $T_a = T_a^{ji}$ the initial start of attack, removing the index ji to ease with notational complexity.

As noted in the analysis of the properties of $\mathcal{O}_i^L$, for all time $t \geq T_a$ the residual r_{ji} can be decomposed into *healthy* and *attacked* components, i.e.

$$r_{ji}(t) = r_{ji}^h(t) + r_{ji}^a(t),$$

where $r_{ji}^h(t)$ has the same definition as $r_{ji}(t)$ in nominal conditions (3.62), independently of T_a , while the attacked component of the residual is given by

$$r_{ji}^a(t) = Z_j \alpha_{ji}(t) + C_j e^{F_j(t-T_a)} H_j \alpha_{ji}(T_a) - C_j \int_{T_a}^t e^{F_j(t-\tau)} \hat{K}_j \alpha_{ji}(\tau) d\tau. \quad (3.65)$$

It is useful to note how both r_{ji}^a and ϵ_{ji}^a (the component of the state estimation error affected

by the attack), can both be seen as outputs of the following LTI system:

$$\begin{cases} \dot{\check{\epsilon}}_{ji} = F_j \check{\epsilon}_{ji} - \widehat{K}_j \alpha_{ji} \\ \epsilon_{ji}^a = \check{\epsilon}_{ji} - H_j \alpha_{ji} \\ r_{ji}^a = C_j \check{\epsilon}_{ji} + Z_j \alpha_{ji} \end{cases} \quad (3.66)$$

where $\check{\epsilon}_{ji} \in \mathbb{R}^{n_j}$ is an internal state. In the remainder of this section, after describing those attacks that are guaranteed to be detected by $\mathcal{O}_{ji}^{UIO}$, we present the classes of attacks that are stealthy to it.

Detectable attacks

Similarly to Proposition 3.2, in the following Proposition we give the condition under which the attacks are *guaranteed* to be detected.

Proposition 3.5. Given γ_{ji}^c transmitted from $\mathcal{S}_j$ to $\mathcal{S}_i$, an attack is guaranteed to be detectable by $\mathcal{O}_{ji}^{UIO}$ at some time $t \leq T_d$ if there exists a time instant $T_d \in (T_a, \infty)$ such that:

$$|r_{ji}^a(T_d)| > 2\bar{r}_{ji}^a \quad (3.67)$$

holds for any one component of $r_{ji}^a(T_d)$. □

Proof. The proof follows that of Proposition 3.2 ■

Stealthy attacks

Let us now focus on the class of attacks which are stealthy to $\mathcal{O}_{ji}^{UIO}$. We highlight analysis on the stealthiness properties of three classes of attacks introduced in Chapter 2:

- i. false-data injection attacks;
- ii. covert attacks;
- iii. replay attacks.

Recalling their definitions in Chapter 2, each of these attacks corresponds to different attack resources available to the attacker. In the following, before introducing conditions for stealthiness, we go over their characteristics once again for completeness.

False-data injection attacks This class of attacks, not requiring any *disclosure* capabilities and performed in open loop, can be implemented by an attacker even if they cannot *eavesdrop* the information being transmitted between controllers. To express it in the distributed scenario considered in this chapter, we consider the attack to be modelled by the function:

$$\alpha_{ji}(t) : \mathbb{R} \rightarrow \mathbb{R}^{p_j}. \quad (3.68)$$

The following Proposition, similar to Proposition 3.3, defines a sufficient condition for a false-data injection attack to be stealthy.

Proposition 3.6. A false-data injection attack $\alpha_{ji}(t) \neq 0$ is stealthy to $\mathcal{O}_{ji}^{UIO}$ if and only if it is a zero-dynamics attack for the LTI system (3.66). $\square$

Proof. By similar arguments to the proof of Proposition 3.3, it is sufficient for the residual affected by the attack to be such that

$$|r_{ji}^a(t)| = 0 \quad \forall t \geq T_a$$

for a false-data injection attack to be stealthy. To achieve this, considering r_{ji}^a as the output of the LTI system (3.66) with input α_{ji} , it is clear that α_{ji} must be output-zeroing. Furthermore, as shown in [PDB13] and summarised in Chapter 2, for any attack on the input channels of a system, it is necessary and sufficient that the attack be a *zero-dynamics attack* for its effect to not be detected from the output. The statement of the proposition follows. $\blacksquare$

Remark 3.15. Recalling that zero-dynamics attacks on LTI systems are equivalent to non-observable inputs [HP98b, PDB13], there are two different types of zero-dynamics attacks that can be performed against system (3.66):

- a. A transmission zero of system $(F_j, -\hat{K}_j, C_j, Z_j)$;

b. An attack such that

$$\alpha_{ji}(t) \in \ker \begin{bmatrix} -\widehat{K}_j \\ Z_j \end{bmatrix}$$

holds for all $t > T_a$.

Note that the dynamics (3.66) rely on matrices internal to $\mathcal{O}_{ji}^{UIO}$: as such an attacker *must* have information on the observer parameters to perform an attack of this type. $\triangleleft$

Replay attacks This class of attacks is capable of recording and then *replaying* actual data transmitted between neighbouring controllers, with a fixed time delay. Specifically, they can be defined as

$$\alpha_{ji}(t) = \gamma_{ji}^c(t - nT) - \gamma_{ji}^c(t) \quad (3.69)$$

where $n \in \mathbb{N}_0$ is a counter, defined as $n \doteq \lceil \frac{t-T_a}{T} \rceil$. As shown in Chapter 2, this class of attacks may be constructed to be stealthy to monitoring schemes under certain conditions, for example when the system against which the attack is leveraged is in steady state. Indeed, replay attacks benefit from exhibiting both the same statistical characteristics of the unaltered signals and the same dynamics of the system. The “strength” of replay attacks, therefore, lies in not requiring any model knowledge and yet being stealthy, under certain conditions.

For a large-scale system composed of subsystems with dynamics as in (3.1), and with monitoring scheme $\mathcal{D}_i$, a sufficient condition for replay attacks to be stealthy to $\mathcal{O}_{ji}^{UIO}$ is presented in the following.

Proposition 3.7. If a replay attack leveraged against γ_{ji}^c received by $\mathcal{S}_i$, equipped with module $\mathcal{O}_{ji}^{UIO}$, is such that:

$$\Sigma_j |\epsilon_{ji}^r(T_a) + H_j v_j(T_a - T) + M_{ji}| \leq \bar{\epsilon}_{ji}(T_a) - |H_j| \bar{v}_j - \Delta \bar{\epsilon}_{ji}(T_a) \quad (3.70)$$

holds, where $\epsilon_{ji}^r(t) \doteq x_j(t - nT) - \hat{x}_{ji}(t)$, and $M_{ji} > 0$, $\Delta \bar{\epsilon}_{ji}(T_a) > 0$ two appropriately defined vectors, then detection test (3.28b) does not hold for any $t \geq T_a$, and the attack is stealthy to $\mathcal{O}_{ji}^{UIO}$. $\square$

Proof. Noting that the replay attack is periodic, we start by introducing the following Lemma, where the Proposition is proved over the first period $[T_a, T_a + T)$. We then extended this result for $t \geq T_a + T$.

Lemma 3.2. If a replay attack is such that

$$\Sigma_j |\epsilon_{ji}^r(T_a) + H_j v_j(T_a - T)| \leq \bar{\epsilon}_{ji}(T_a) - |H_j| \bar{v}_j \quad (3.71)$$

then detection test (3.28b) is not satisfied for any $t \in [T_a, T_a + T)$. $\square$

Proof. Over the first period of the replay attack $t \in [T_a, T_a + T)$, the UIO residual error $r_{ji}(t)$ takes the form:

$$\begin{aligned} r_{ji}(t) &= \gamma_{ji}^c(t) - \hat{\gamma}_{ji}(t) \\ &= y_j(t - T) - C_j \hat{x}_{ji}(t) \\ &= C_j \epsilon_{ji}^r + v_j(t - T) \end{aligned} \quad (3.72)$$

with $\epsilon_{ji}^r(t)$ following the dynamics

$$\dot{\epsilon}_{ji}^r(t) = F_j \epsilon_{ji}^r(t) + w_j(t - T) - \tilde{K}_j v_j(t - T) - H_j \dot{v}_j(t - T), \quad (3.73)$$

for $t \in [T_a, T_a + t)$. The solution of (3.73) can be derived, through integration by parts, as:

$$\begin{aligned} \epsilon_{ji}^r(t) &= e^{F_j(t-T_a)} (\epsilon_{ji}^r(T_a) + H_j v_j(T_a - T)) - H_j v_j(t - T) \\ &\quad + \int_{T_a}^t e^{F_j(t-\tau)} [S_j w_j(\tau - T) - \hat{K}_j v_j(\tau - T)] d\tau. \end{aligned} \quad (3.74)$$

Recalling the definition of the estimation error bound $\bar{\epsilon}_{ji}$ in (3.64), it can be rewritten as:

$$\bar{\epsilon}_{ji}(t) = e^{-\sigma_j(t-T_a)} \Sigma_j [\bar{\epsilon}_{ji}(T_a) - |H_j| \bar{v}_j] + |H_j| \bar{v}_j + \int_{T_a}^t e^{-\sigma_j(t-\tau)} \Sigma_j [|S_j| \bar{w}_j + |\hat{K}_j| \bar{v}_j] d\tau. \quad (3.75)$$

To achieve stealthiness over the first period of the attack, it is sufficient that $|\epsilon_{ji}^r(t)| \leq \bar{\epsilon}_{ji}(t)$, for all $t \in [T_a, T_a + T)$. By directly comparing (3.73) and (3.75), all terms in (3.73) are guaranteed

to be bounded by their equivalent in (3.75), except for

$$e^{F_j(t-T_a)} (\epsilon_{ji}^r(T_a) + H_j v_j(T_a - T)), \quad (3.76)$$

given the definition of the disturbance bounds (3.3) in Assumption 3.1. Hence, since

$$|e^{F_j(t-T_a)} (\epsilon_{ji}^r(T_a) + H_j v_j(T_a - T))| \leq e^{-\sigma_j(t-T_a)} \Sigma_j |(\epsilon_{ji}^r(T_a) + H_j v_j(T_a - T))| \quad (3.77)$$

holds by construction of σ_j and Σ_j , (3.71) is a sufficient condition for stealthiness to be guaranteed for all $t \in (T_a, T_a + T)$.

To prove that (3.71) is sufficient for $|\epsilon_{ji}^r(T_a)| \leq \bar{\epsilon}_{ji}(T_a)$, we exploit the property of Σ_j such that $\Sigma_j \geq I$. Thus, by applying the inverse triangle inequality, we finally show that:

$$\Sigma_j |\epsilon_{ji}^r(T_a) + H_j v_j(T_a - T)| \leq |\epsilon_{ji}^r(T_a) + H_j v_j(T_a - T)| \leq |\epsilon_{ji}^r(T_a)| - |H_j| \bar{v}_j. \quad (3.78)$$

This completes the proof of Lemma 3.2. ■

We now extend the result of Lemma 3.2 for all $t \geq T_a$, thus proving the Proposition. Let us start by showing that we can define similar conditions to (3.71) in Lemma 3.2, for all time instances $\tau \in \mathcal{T} \doteq \{\tau \in \mathbb{R} | \tau = T_a + kT, k \in \mathbb{N}\}$, i.e. that satisfaction of the inequality

$$\Sigma_j |\epsilon_{ji}^r(\tau) + H_j v_j(T_a - T)| \leq \bar{\epsilon}_{ji}(\tau) - |H_j| \bar{v}_j, \quad (3.79)$$

is a sufficient condition on $\epsilon_{ji}^r(\tau)$ such that $|\epsilon_{ji}^r(t)| \leq \bar{\epsilon}_{ji}(t)$ for all $t \in [\tau, \tau + T)$. Hence, to complete the proof, we must show that (3.70) implies (3.79). To do this, it is necessary to characterise the solution of $\epsilon_{ji}^r(\tau), \tau \in \mathcal{T}$, as $k \rightarrow \infty$. We start by noting that, for $\tau \in \mathcal{T}$, the state estimation while under attack is:

$$\begin{aligned} \hat{x}_{ji}(\tau) &= z_{ji}(\tau) + H_j \gamma_{ji}^c(\tau) \\ &= e^{F_j k T} z_{ji}(T_a) + H_j \gamma_{ji}^c(\tau) + \int_{T_a}^{\tau} e^{F_j(T_a + kT - s)} \hat{K}_j \gamma_{ji}^c(s) ds \\ &= e^{F_j k T} z_{ji}(T_a) + H_j \gamma_{ji}^c(T_a - T) + \sum_{\kappa=0}^{k-1} e^{\kappa T F_j} \int_{T_a - T}^{T_a} e^{F_j(T_a - s)} \hat{K}_j \gamma_{ji}^c(s) ds. \end{aligned} \quad (3.80)$$

As such, $\epsilon_{ji}^r(\tau)$ can be expressed as

$$\epsilon_{ji}^r(\tau) = \epsilon_{ji}^r(T_a) + \underbrace{\left(I - e^{F_j k T} \right) z_{ji}(T_a) - \sum_{\kappa=0}^{k-1} e^{\kappa T F_j} \int_{T_a-T}^{T_a} e^{F_j(T_a-s)} \widehat{K}_j \gamma_{ji}^c(s) ds}_{\Delta \epsilon_{ji}(k)}. \quad (3.81)$$

Given the Hurwitz stability of F_j , the series in (3.81),

$$\sum_{\kappa=0}^{k-1} e^{\kappa T F_j} \int_{T_a-T}^{T_a} e^{F_j(T_a-s)} \widehat{K}_j \gamma_{ji}^c(s) ds,$$

is convergent. Therefore, also $\Delta \epsilon_{ji}(k)$ defined in (3.81) is convergent. Consequently, there exists $M_{ji} \geq 0$ such that, for $\tau \in \mathcal{T}$,

$$|\epsilon_{ji}^r(\tau)| \leq |\epsilon_{ji}^r(T_a) + M_{ji}|$$

holds. Specifically, $M_{ji} \geq 0$ is defined to satisfy

$$|\Delta \epsilon_{ji}(k)| \leq M_{ji}, \forall k \in \mathbb{N}.$$

Let us define $M_{ji} \doteq \sup_{k \in \mathbb{N}} |\Delta_{ji}(k)|$. Given that M_{ji} is an upper bound of $|\Delta \epsilon_{ji}(k)|$ for all $k \in \mathbb{N}$ by definition, the following holds:

$$|\epsilon_{ji}^r(T_a + kT) + H_j v_j(T_a - T)| \leq |\epsilon_{ji}^r(T_a) + M_{ji} + H_j v_j(T_a - T)|, \quad \forall k \in \mathbb{N}_0. \quad (3.82)$$

Therefore, to complete the proof, it is sufficient to compare the right hand side of (3.82) with the bound $\bar{\epsilon}_{ji}(t)$ as $t \rightarrow \infty$. Note that $\bar{\epsilon}_{ji}(t)$ is monotonic by design. Let us define $\Delta_{ji}(T_a)$ to account for whether $\bar{\epsilon}_{ji}$ is monotonically increasing or decreasing:

$$\Delta \epsilon_{ji}(T_a) \doteq \max \left(0, \lim_{t \rightarrow \infty} \bar{\epsilon}_{ji}(T_a) - \bar{\epsilon}_{ji}(t) \right). \quad (3.83)$$

Hence, the bound on the estimation error satisfies:

$$\bar{\epsilon}_{ji}(T_a + kT) \geq \bar{\epsilon}_{ji}(T_a) - \Delta \bar{\epsilon}_{ji}(T_a), \forall k \in \mathbb{N}. \quad (3.84)$$

Thus, (3.70) can be shown to be a sufficient condition for (3.79). Indeed, recall that $\Sigma_j \geq I$ holds by Lemma 3.1. Hence, if (3.70) holds, (3.71) also holds, as

$$\begin{aligned} \Sigma_j \left| \epsilon_{ji}^r(T_a + kT) + H_j v_j(T_a - T) \right| &\leq \Sigma_j \left| \epsilon_{ji}^r(T_a) + M_{j,i} + H_j v_j(T_a - T) \right| \\ &\leq \bar{\epsilon}_{ji}(T_a) - \Delta \bar{\epsilon}_{ji}(T_a) - |H_j| \bar{v}_j \\ &\leq \bar{\epsilon}_{ji}(T_a + kT) - |H_j| \bar{v}_j, \forall k \in \mathbb{N}. \end{aligned}$$

Consequently, $|\epsilon_{ji}^r(t)| \leq \bar{\epsilon}_{ji}(t)$ holds for all $t \geq T_a$. ■

Although condition (3.70) offers a sufficient condition for stealthiness of a replay attack, it depends on variables that are unknown to the attacker, namely the estimation error and noise. As such, it is not possible for an attacking agent to *guarantee* that (3.70) be satisfied. However, as long as the attacker selects T_a and T appropriately, i.e. as long as $\gamma_{ji}^c(T_a) \approx \gamma_{ji}^c(T_a - T)$, it is *likely*¹¹ that (3.28b) does not hold for all $t \geq T_a$, and therefore that the attacker is stealthy.

Covert attacks This is the final class of attacks considered. When performing covert attacks, attackers are not only able to eavesdrop the messages transmitted over the communication network, but also have sufficient knowledge of the model of $\mathcal{S}_j$ to *simulate* a synthetic state which is consistent with the real trajectory of state x_j . As such the attack is defined as:

$$\alpha_{ji}(t) = \gamma_{ji}^a(t) - \gamma_{ji}^c(t) \quad (3.85)$$

with $\gamma_{ji}^a(t) \in \mathbb{R}^{p_j}$ defined as the output of the following LTI system:

$$\mathcal{S}_{ji}^a : \begin{cases} x_{ji}^a = A_{jj} x_{ji}^a + \bar{E}_j \bar{d}_{ji}^a \\ \gamma_{ji}^a = C_j x_{ji}^a \end{cases}, \quad x_{ji}^a(T_a) = C_j^\dagger \gamma_{ji}^c(T_a) \quad (3.86)$$

with $\bar{d}_{ji}^a \in \mathbb{R}^{\bar{q}_j}$ defined by the attacker to achieve its own objectives, and initial state $x_{ji}^a(T_a)$. Clearly, for $\text{rank } C_j < n_j$, the attacker may introduce a nonzero initial state error $\Delta x_{ji}^a(T_a) \doteq x_j(T_a) - x_{ji}^a(T_a) \in \ker C_j$. Given detectability of (A_{jj}, C_j) from Assumption 3.2, however, any

¹¹Here we abuse the term “likely”, as “likelihood” cannot be computed – recall that both w_i and v_i are unstructured for all $i \in \mathcal{N}$.

error in initial state is either observable from $\gamma_{ji}^a(t), t > T_a$ or decays to zero exponentially. Thus, if this deviation is sufficiently large, it may be such that it satisfies (3.28b), making it possible for $\mathcal{O}_{ji}^{UIO}$ to detect it.

Assumption 3.11. *Initial state of covert attacker's internal state $x_{ji}^a(T_a)$ does not introduce any initial state error, i.e. $\Delta x_{ji}^a(T_a) = 0$.* $\triangleleft$

Remark 3.16. Note that, from the point of view of *defending* $\mathcal{S}_i$, Assumption 3.11 is a worst case scenario, as it gives a larger advantage to the attacking agent. $\triangleleft$

Proposition 3.8. Consider a covert attack as in (3.85), with γ_{ji}^a the output of LTI system (3.86), on the communication received by $\mathcal{S}_i$ from $\mathcal{S}_j$; suppose $\mathcal{S}_i$ is monitored by module $\mathcal{O}_{ji}^{UIO}$ defined in (3.57). If $v_j(T_a)$ is such that

$$C_j \Sigma_j \left| \left(C_j^\dagger - H_j \right) v_j \right| \leq |Z_j| \bar{v}_j \quad (3.87)$$

the attack is stealthy for all $t \geq T_a$, as (3.28b) is guaranteed not to hold.

Proof. For $t = T_a$, the residual is

$$r_{ji}(T_a) = C_j x_{ji}^a - \hat{\gamma}_{ji} = C_j (x_j(T_a) - \hat{x}_{ji}(T_a))$$

and as such, given Proposition 3.4, (3.28b) does not hold.

For $t \geq T_a$, defining $\varepsilon_{ji}^a(t) \doteq x_{ji}^a - \hat{x}_{ji}(t)$, with dynamics¹²

$$\dot{\varepsilon}_{ji}^a = F_j \varepsilon_{ji}^a + S_j \bar{E}_j \bar{d}_{ji}^a = F_j \varepsilon_{ji}^a,$$

the residual error $r_{ji}(t) = C_j \varepsilon_{ji}^a(t) = C_j e^{F_j t} \varepsilon_{ji}^a(T_a)$ is

$$\begin{aligned} r_{ji}(t) &= C_j e^{F_j(t-T_a)} [\varepsilon_{ji}(0) + H_j v_j(0)] + C_j e^{F_j(t-T_a)} \left(C_j^\dagger - H_j \right) v_j(T_a) \\ &\quad + C_j \int_0^{T_a} e^{F_j(t-\tau)} \left[S_j w_j(\tau) - \hat{K}_j v_j(\tau) \right] d\tau. \end{aligned}$$

¹²We introduce ε_{ji}^a to avoid ambiguity with ϵ_{ji}^a , the component of the estimation error $\epsilon_{ji} = x_j - \hat{x}_{ji}$ affected by the attack.

By direct comparison between this and the definition of $\bar{r}_{ji}$ in (3.60), the only term not guaranteed to be bounded is

$$C_j e^{F_j(t-T_a)} (C_j^\dagger - H_j) v_j(T_a). \quad (3.88)$$

Hence, it is sufficient for

$$\left| C_j e^{F_j(Tt-T_a)} (C_j^\dagger - H_j) v_j(T_a) \right| \leq |Z_j| \bar{v}_j + C_j \int_{T_a}^t e^{-\sigma_j(t-\tau)} \Sigma_j \left[|S_j| \bar{w}_j + |\hat{K}_j| \bar{v}_j \right] d\tau \quad (3.89)$$

to be satisfied to guarantee that (3.28b) does not hold for any $t \geq T_a$. By recalling that $|e^{F_j t}| \leq \Sigma_j e^{-\sigma_j t}$ by construction, it is sufficient for (3.87) to hold for a covert attack as in (3.85) with dynamics (3.86) to be stealthy for monitoring module $\mathcal{O}_{ji}^{UIO}$ for all $t \geq T_a$. ■

Remark 3.17. Note that, given the scenario we are considering, we have assumed for both covert and replay attacks that the malicious agents do not have access to any actuation channels, differently from what is normally considered in the literature. Indeed, recalling Figure 2.4 and discussion in Chapter 2, in the literature malicious agents deploying covert or replay attacks have the disruption resources required to inject attacks on the actuation channels of a plant.

Here, however, given Assumption 3.7, the behaviour of the plant cannot be altered directly by the attacker. Hence, in this scenario, for the malicious agent to have an effect on the overall system, they must rely on the state of $\mathcal{S}_{ji}^a$ being different to the state of $\mathcal{S}_j$, in the case of covert attacks, or from its behaviour during the “recording” phase of the replay attack. This can be seen as the attacker trying to deceive $\mathcal{C}_i$ about what the real state of $\mathcal{S}_j$ is (covert attack), or to obfuscate any changes in $\bar{d}_j$ (replay attacks). ◁

Having presented the properties of both modules $\mathcal{O}_i^L$ and $\mathcal{O}_{ji}^{UIO}$ within $\mathcal{D}_i$, we are now interested in analysing how their simultaneous use in $\mathcal{D}_i$ influences the diagnoser’s performance.

3.6 Detectability properties of $\mathcal{D}_i$

We now analyze the properties of $\mathcal{D}_i$ as a whole. Given its design and operation (recall Algorithm 1), its properties are the combination of both $\mathcal{O}_i^L$ and $\mathcal{O}_{ji}^{UIO}$. Specifically, recall from

(3.28a)-(3.28b) that it is sufficient for either

$$|r_i| > \bar{r}_i \quad \text{or} \quad |r_{ji}| > \bar{r}_{ji}$$

to hold for an attack to be detected by $\mathcal{D}_i$.

In the following, by combining Propositions 3.2 and 3.5, giving sufficient conditions for detection, we characterise those attacks which are guaranteed to be detected by $\mathcal{D}_i$.

Proposition 3.9. Given a subsystem $\mathcal{S}_i$ equipped with controller $\mathcal{C}_i$ and detector $\mathcal{D}_i$, consider from some time $T_a^{ij} > 0$ an attack $\alpha_{ji} \neq 0$ is active on the signal γ_{ji}^c received from $\mathcal{C}_j$. If the cyber-attack is such that, at some time $T_d^{ij} > T_a^{ij}$:

$$|r_i^a(T_d^{ij})| > 2\bar{r}_i(T_d^{ij}) \quad \text{or} \quad |r_{ji}^a(T_d^{ij})| > 2\bar{r}_{ji}(T_d^{ij}) \quad (3.90)$$

then $\mathcal{D}_i$ is guaranteed to detect $\alpha_{ji}(t)$ at some time $t \leq T_d^{ij}$. $\square$

Proposition 3.9 demonstrates how the combination of the two modules $\mathcal{O}_i^L$ and $\mathcal{O}_{ji}^{UIO}$ within $\mathcal{D}_i$ improve its detection properties. However, it is through the analysis of those attacks that are stealthy to $\mathcal{D}_i$ that it becomes evident how the combination of $\mathcal{O}_i^L$ and $\mathcal{O}_{ji}^{UIO}$ improves the overall performance of $\mathcal{D}_i$. Indeed, except in some pathological scenarios, detection occurs even when (3.90) is not satisfied.

Proposition 3.10. Given a subsystem $\mathcal{S}_i$ equipped with detector $\mathcal{D}_i$, consider from some time $T_a^{ij} > 0$ an attack $\alpha_{ji} \neq 0$ is active on the received information γ_{ji}^c . If the cyber-attack function is such that, for all $t \geq T_a^{ji}$:

$$|r_i^a(t)| = 0 \quad \text{and} \quad |r_{ji}^a(t)| = 0 \quad (3.91)$$

then $\alpha_{ji}(\cdot)$ is stealthy to detector $\mathcal{D}_i$. $\square$

In a similar spirit to Proposition 3.9, to be stealthy to $\mathcal{D}_i$, an attack function $\alpha_{ji}(\cdot)$ must be stealthy to both detection modules. Recall from Proposition 3.6 that any false-data injection attack on γ_{ji}^c must be a zero-dynamics attack for dynamics (3.66) to be stealthy. Given

Remark 3.15, to be stealthy to $\mathcal{D}_i$ the attack signal $\alpha_{ji}, j \in \widehat{\mathcal{N}}_i(t)$ must either be such that

$$\alpha_i(t) \in \ker \left(\begin{bmatrix} -\widehat{\mathbb{K}}_i(t) \\ \mathbb{Z}_i(t) \end{bmatrix} \right) \cup \ker(\mathbb{A}_i(t))$$

where $\mathbb{A}_i(t) = \text{row}_{j \in \widehat{\mathcal{N}}_i(t)} (A_{ij})$, $\widehat{\mathbb{K}}_i(t) \doteq \text{diag}_{j \in \widehat{\mathcal{N}}_i(t)} (\widehat{K}_j)$ and $\mathbb{Z}_i(t) \doteq \text{diag}_{j \in \widehat{\mathcal{N}}_i(t)} (Z_j)$, while also satisfying $\beta_{ji}(0)\alpha_{ji}(T_a^{ji}) = 0, \forall j \in \widehat{\mathcal{N}}_i(t)$, or it must simultaneously be a transmission zero of $-C_j(sI - F_j)^{-1}\widehat{K}_j$ and lie within $\ker(\mathbb{A}_{ij}(t))$.

Having seen both detectable and stealthy attacks to $\mathcal{D}_i$, we now focus our attention to covert and replay attacks. Specifically, we analyse the case where they might be detected by $\mathcal{O}_i^L$, while being stealthy to $\mathcal{O}_{ji}^{UIO}$, given conditions in Propositions 3.7 and 3.8. This is done to highlight how the integration of the two modules within $\mathcal{D}_i$ allows us to perform detection even of very sophisticated attacks, such as covert and replay attacks.

Given that covert and replay attacks present similar structure, we start by presenting a brief summary of their definition and behaviour, showing how both replay and covert attacks exploit the same dynamic characteristics of $\mathcal{S}_j$. Recall that both can be described as the attack function:

$$\alpha_{ji}(t) = -\gamma_{ji}^c(t) + \gamma_{ji}^a(t) \quad (3.92)$$

with $\gamma_{ji}^a(t)$ the output of LTI system

$$\mathcal{S}_{ji}^a : \begin{cases} \dot{x}_{ji}^a = A_{jj}x_{ji}^a + \bar{E}_j\bar{d}_{ji}^a + w_{ji}^a \\ \gamma_{ji}^a = C_jx_{ji}^a + v_{ji}^a \end{cases} \quad (3.93)$$

where the values of $\bar{d}_{ji}^a$, w_{ji}^a and v_{ji}^a are given in Table 3.3. For replay attacks they are *implicitly* defined by the attacker's decision of the "replay period" T , whereas for covert attacks they are *explicitly* defined by the attacker itself. Note that the definition of the dynamics in (3.93) abuses notation slightly for replay attacks: indeed, given periodic behavior of $x_{ji}^a(t)$, it may be discontinuous for time $t \in \mathcal{T} \doteq \{n \in \mathbb{N}_0 | t = T_a + nT\}$, and as such (3.93) holds for all $t \geq T_a^{ij}, t \notin \mathcal{T}$.

Table 3.3: Interpretation of replay and covert attacks as LTI system (3.93)

	Replay attacks	Covert attacks
$x_{ji}^a(T_a)$	$x_j^a(T_a - T)$	$C_j^\dagger \gamma_{ji}^c(T_a)$
$\bar{d}_{ji}^a(t)$	$\bar{d}_j(t - nT)$	$\bar{d}_{ji}^a(t)$
$w_{ji}^a(t)$	$w_j(t - nT)$	0
$v_{ji}^a(t)$	$v_j(t - nT)$	0

For both replay and covert attacks, it is possible to rewrite

$$\alpha_{ji}(t) = \Delta \gamma_{ji}^a(t) \quad (3.94)$$

with $\Delta \gamma_{ji}^a(t)$ the output of the LTI system

$$\Delta \mathcal{S}_{ji}^a : \begin{cases} \Delta \dot{x}_{ji}^a = A_{jj} \Delta x_{ji}^a + \bar{E}_j \Delta \bar{d}_{ji}^a + \Delta w_{ji}^a \\ \Delta \gamma_{ji}^a = C_j \Delta x_{ji}^a + \Delta v_{ji}^a \end{cases} \quad (3.95)$$

with $\Delta \bar{d}_{ji}^a, \Delta w_{ji}^a$ and Δv_{ji}^a easily obtainable from Table 3.3, and $x_{ji}^a = x_j + \Delta x_{ji}^a$. The explicit solution $\Delta x_{ji}^a(t)$ to (3.95) is:

$$\Delta x_{ji}^a(t) = e^{A_{jj}(t-T_a^{ij}-nT)} \Delta x_{ji}^a(T_a^{ij} + nT) + \int_{T_a^{ij}+nT}^t e^{A_{jj}(t-\tau)} [\bar{E}_j \Delta \bar{d}_{ji}^a(\tau) + \Delta w_{ji}^a(\tau)] d\tau \quad (3.96)$$

where, in case of covert attacks, $nT \doteq 0$.

Theorem 3.2. Given a subsystem $\mathcal{S}_i$ monitored by $\mathcal{D}_i$, consider a covert or replay attack $\alpha_{ji}(t)$ as in (3.94), the output of LTI system (3.95), stealthy to $\mathcal{O}_{ji}^{UIO}$ in $\mathcal{D}_i$. If, for some time $T_d \geq \tilde{T}_a$, $\tilde{T}_a \doteq \min_{j \in \mathcal{N}_i} T_a^{ji}$, Δx_{ji}^a is such that:

$$\left| C_i \int_{\tilde{T}_a}^{T_d} e^{A_{Li}(t-\tau)} \left[\sum_{j \in \hat{\mathcal{N}}_i(t)} A_{ij} \Delta x_{ji}^a(\tau) \right] d\tau \right| \geq 2\bar{r}_i(T_d), \quad (3.97)$$

then the detection test of $\mathcal{O}_i^L$ (3.28a) is guaranteed to hold for some $t \leq T_d$, thus achieving detection. $\square$

Proof. To prove detection occurs in $\mathcal{D}_i$, either (3.28a) or (3.28b) must hold for some $t > \tilde{T}_a$.

Given assumption that $\alpha_{ji}(t)$ is either a covert or a replay attack, it is stealthy to $\mathcal{O}_{ji}^{UIO}$. As such, we are interested in the detection properties of $\mathcal{O}_i^L$.

Given the definition of $\alpha_{ji}(t)$ in (3.94), and recalling Remark 3.11 and Assumption 3.3, the attacked portion of the residual defined in (3.47) can be rewritten as:

$$r_i^a(t) = C_i \int_{T_a^{ij}}^{T_d} e^{A_{Li}(t-\tau)} \left[\sum_{j \in \hat{\mathcal{N}}_i(t)} A_{ij} \Delta x_{ji}^a(\tau) \right] d\tau, \quad (3.98)$$

noting that $|\Delta v_{ji}^a(t)| \leq \bar{v}_j(t)$ holds for all $t \geq T_a^{ij}$. The rest of the proof follows that of Proposition 3.2. $\blacksquare$

Theorem 3.2 provides an interesting insight in the characteristics of $\mathcal{D}_i$ as a whole, and of the interplay between $\mathcal{O}_i^L$ and $\mathcal{O}_{ji}^{UIO}$. Indeed, it proves that it is possible, and in certain conditions *guaranteed*, to detect covert and replay attacks, stealthy to $\mathcal{O}_{ji}^{UIO}$, through the local detector $\mathcal{O}_i^L$.

This is critical, as Theorem 3.2 gives bounds for *how much* a malicious agent on communication line (j, i) implementing a covert or a replay attack can alter the information received by $\mathcal{S}_i$, and therefore the overall behavior of the large-scale system, by establishing the maximum deviation Δx_{ji}^a from x_j before being detected by $\mathcal{O}_i^L$.

In turn this implies that covert attacks cannot choose $\Delta \bar{d}_{ji}^a$ arbitrarily if they want to remain stealthy, but rather $\Delta \bar{d}_{ji}^a$ must remain *small*. On the other hand, Theorem 3.2 suggests that if the operating conditions of $\mathcal{S}_j$ vary significantly over time, then replay attacks are guaranteed to be detected.

Remark 3.18. The above properties of $\mathcal{D}_i$ hold even if Assumption 3.3 does not hold, i.e. even if an estimate of $\xi_i = \sum_{j \in \mathcal{N}_i} A_{ij} x_j$ is not directly recoverable from γ_{ji}^c , and therefore it is necessary to compute the dynamics of $\hat{x}_i$ as in Section 3.4.3. Indeed, although

$$|r_{ji}| \leq \bar{r}_{ji} \quad (3.99)$$

holds for covert and replay attacks, this is due to the fact that the attack is able to have the estimate $\hat{x}_{ji}$ follow the attacked state Δx_{ji}^a . Thus $\varepsilon_{ji}^a \doteq \Delta x_{ji}^a - \hat{x}_{ji}$ always satisfies $|\varepsilon_{ji}^a| \leq \bar{\varepsilon}_{ji}$.

However, given that $\mathcal{O}_i^L$ relies on the physical coupling between subsystems to detect attacks, the estimation error ϵ_i is not proportional to ε_{ji}^a , but to $\epsilon_{ji} = x_j - \hat{x}_{ji}$ itself.

Therefore, $\mathcal{O}_i^L$ is still capable of detecting attacks, even when covert or replay attacks are levelled against the diagnoser $\mathcal{D}_i$. Furthermore, $\mathcal{O}_i^L$ is guaranteed to detect the attack when

$$\left| C_i \int_{\tilde{T}_a}^{T_d} e^{A_{Li}(t-\tau)} \left[\sum_{j \in \hat{\mathcal{N}}_i(t)} A_{ij} \epsilon_{ji}(\tau) \right] d\tau \right| \geq 2\bar{r}_i(T_d) \quad (3.100)$$

holds. Clearly, inequality (3.100) is equivalent to (3.97), having substituted Δx_{ji}^a with ϵ_{ji} . $\triangleleft$

3.7 Isolability of attacks in the communication network

Having analyzed in detail the detection properties of $\mathcal{D}_i$, we now focus on isolation. As introduced in Chapter 2, and discussed in [WS19], cyber-attack isolation is similar to fault isolation. In light of the problem we have discussed thus far, we consider the following:

Problem 3.2 (Distributed Attack Isolation). Given a large-scale system divided into subsystems $\mathcal{S}_i$ with dynamics as in (3.1), receiving information γ_{ji}^c from their neighbours $\mathcal{S}_j$, $j \in \mathcal{N}_i$, defined in (3.7), the problem of *distributed attack isolation* is equivalent to identifying the communication link (j, i) over which attack $\alpha_{ji}(t) \neq 0$. $\triangleleft$

Given the architecture we have presented in this chapter, Problem 3.2 can be resolved differently, depending on whether $\mathcal{O}_{ji}^{UIO}$ detects the attack or not. In the remainder of this section, we analyse different scenarios leading to isolation.

3.7.1 Isolation by $\mathcal{O}_{ji}^{UIO}$

Let us start by noting that once $\mathcal{O}_{ji}^{UIO}$ detects an attack, this attack is also isolated.

Proposition 3.11. Consider $\mathcal{S}_i$ receiving communication signals γ_{ji}^c from its neighbours and monitored by distributed diagnoser $\mathcal{D}_i$. If an attack $\alpha_{ji}(t) \neq 0$ is such that:

$$|r_{ji}(t)| > \bar{r}_{ji}(t) \quad (3.101)$$

then the attack is isolated to the communication link between $\mathcal{C}_j$ to $\mathcal{C}_i$. $\square$

Proposition 3.11 is a direct consequence of the construction of $\mathcal{O}_{ji}^{UIO}$. Indeed, recalling the dynamics of the UIO in (3.57), neither the internal UIO state z_{ji} nor the state estimate $\hat{x}_{ji}$ depend on any $\gamma_{ki}^c, k \in \mathcal{N}_i \setminus \{j\}$, and thus is not influenced by α_{ki} . Therefore, once (3.28b) is satisfied for $\mathcal{O}_{ji}^{UIO}$, attack $\alpha_{ji}(t)$ is isolated to communication link (j, i) , solving Problem 3.2.

3.7.2 Isolation by $\mathcal{O}_i^L$

Separate, and more complex, is the case where an attack is detected by $\mathcal{O}_i^L$ and not by $\mathcal{O}_{ji}^{UIO}$. We focus on the cases where an attack is stealthy to $\mathcal{O}_{ji}^{UIO}$. Indeed, if this is not the case, there may be a time $T_d^{UIO} > T_d^L$ for which $\mathcal{O}_{ji}^{UIO}$ detects, and thus isolates, the attack α_{ji} , with T_d^L the time of detection by $\mathcal{O}_i^L$. Specifically, let us suppose the presence of either a covert or replay attack on γ_{ji}^c .

The isolation properties of $\mathcal{D}_i$ in this case depend on the properties of the physical interconnection matrices A_{ij} , as any isolation scheme must rely on the error $\xi_i - \hat{\xi}_i$ to achieve isolation. We start by analysing the case in which the interconnection matrices A_{ij} do not share the same column space.

Proposition 3.12. Suppose that, for any $j, k \in \mathcal{N}_i, j \neq k$,

$$\text{Im}[A_{ij}] \cap \text{Im}[A_{ik}] = \emptyset. \quad (3.102)$$

Then, Problem 3.2 can be solved by input reconstruction of Δx_{ji}^a from r_i . $\square$

Proof. Recall that under covert or replay attack, r_i can be seen as the output of the LTI system:

$$\begin{aligned} \dot{\epsilon}_i &= A_{Li}\epsilon_i - \sum_{j \in \mathcal{N}_i} A_{ij}v_j - \sum_{i \in \mathcal{N}} A_{ij}\Delta x_{ji}^a + w_i - L_i v_i \\ r_i &= C_i\epsilon_i + v_i \end{aligned} \quad (3.103)$$

and that

$$\text{rank}(C_i A_{ij}) = \text{rank}(A_{ij}), \quad \forall j \in \mathcal{N}_i$$

from Assumption 3.10. A further consequence of Assumption 3.10 is that there are no unstable transmission zeros of the transfer function

$$G(s) = C_i(sI - A_{ii})^{-1}\bar{A}_{ij}, \quad (3.104)$$

for any $j \in \mathcal{N}_i$, where $\bar{A}_{ij}$ is a full-column rank matrix¹³ representing a basis of $\text{Im}[A_{ij}]$. Thus, for appropriate design of L_i in (3.29), there are no unstable zero dynamics for the transfer function

$$G_r^{ij}(s) = C_i(sI - A_{Li})^{-1}\bar{A}_{ij}. \quad (3.105)$$

Finally, given (3.102), $G_r^{ij}(s) \neq G_r^{ik}(s)$, for any $j, k \in \mathcal{N}_i, j \neq k$, and it is therefore possible to estimate $\overline{\Delta x_{ji}^a} \neq 0$, i.e. the component of $\Delta x_{ji}^a \notin \ker A_{ij}$ satisfying $\bar{A}_{ij}\overline{\Delta x_{ji}^a} = A_{ij}\Delta x_{ij}^a$, through unknown-input reconstruction techniques [CT98, BFP09]. Thus Problem 3.2 is solved, and the attack is isolated. ■

Although Proposition 3.12 shows a condition under which it is possible to isolate attacks from detection in $\mathcal{O}_i^L$, it is worth noting that for a large class of LSSs, (3.102) may not be satisfied. Indeed, if LSS is composed of a set of structurally similar subsystems, input reconstruction techniques fail to estimate $\Delta x_{ji}^a \neq \Delta x_{ki}^a$ from r_i , which is the case in several applications of interest, such as power networks (see, for example, [RFFT13]) or DC microgrids [TRFT17]. We provide a brief example, based on the application to DC microgrids, for clarity.

Example 3.2. Recall, from the discussion on DC microgrids in Chapter 2, that all matrices A_{ij} are defined as the following

$$A_{ij} = \begin{bmatrix} \frac{R_{ij}}{C_{ti}} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad (3.106)$$

where R_{ij} is the resistance of the power line connecting DGU i to DGU j , and C_{ti} is the i -th filter capacitance. Furthermore, recall the definition of $G_r^{ij}(s)$ in (3.105), with $\bar{A}_{ij}$ defined such that its columns are a full column rank basis for $\text{Im}(A_{ij})$.

¹³For further discussions on input observability, see Appendix A and specifically the remarks following Theorem A.5.

It is clear to see that, for all $j \in \mathcal{N}_i, i \in \mathcal{N}$, $\bar{A}_{ij}$ is appropriately defined as:

$$\bar{A}_{ij} \doteq \begin{bmatrix} * & 0 & 0 \end{bmatrix}^\top, \quad (3.107)$$

with free parameter $*$. As such, it is possible to define matrices $\bar{A}_{ij}$ such that

$$G_r^{ij}(s) = G_r^{ik}(s), \quad \forall j, k \in \mathcal{N}_i, j \neq k. \quad (3.108)$$

Hence, any attack Δx_{ji}^a is indistinguishable from a corresponding attack Δx_{ki}^a , although rescaled.

▽

We must therefore find a different solution to Problem 3.2. For the remainder of this section, let us suppose that there is only one attack active for time $t \geq T_a$:

Assumption 3.12. For $\mathcal{S}_i$ receiving attacked signals γ_{ji}^c , $\hat{\mathcal{N}}_i(t) = \{j\}$, for all $t \geq T_a$. ◁

Note that, if (3.102) does not hold, then it is not possible to uniquely reconstruct Δx_{ji}^a , with unknown $j = \hat{\mathcal{N}}_i$, despite the fact that $\hat{\delta}_{ji}$ may be computed¹⁴, where $\hat{\delta}_{ji}$ is an estimate of

$$\delta_{ji} = A_{ij} \Delta x_{ji}^a, \quad (3.109)$$

the effect of the attack on the residual r_i . Indeed, to compute an estimate of Δx_{ji}^a , $\bar{A}_{ij}$ must be known, with $\bar{A}_{ij}$ a full column rank matrix whose columns form a basis for $\text{Im} A_{ij}$.

However, since we are considering Δx_{ji}^a to be the result of a covert or replay attack on γ_{ji}^c , its trajectory satisfies the dynamics in (3.95), that are known to $\mathcal{D}_i$ [Table 3.2]. Therefore, we may attempt to estimate Δx_{ji}^a using $\hat{\delta}_{ji}$ as an “output” of the LTI system (3.95). Again, since $\hat{\mathcal{N}}_i$ is unknown to $\mathcal{D}_i$, Δx_{ki}^a must be estimated for each $k \in \mathcal{N}_i$. This can be seen as similar to a generalised observer scheme [Cla78], where N_i possible estimates of Δx_{ji}^a are computed, and all but one are discarded, isolating the communication line $(j, i), j \in \hat{\mathcal{N}}_i$.

Given that Δx_{ji}^a is subject to external disturbances, namely Δd_{ji}^a , any observer estimating Δx_{ji}^a must be decoupled to them. Thus, the conditions in Assumption 3.10 must hold, with C_j

¹⁴The existence of an input estimator to compute $\hat{\delta}_{ji}$ is guaranteed by Assumption 3.10.

substituted by A_{ij} , for all $i \in \mathcal{N}_j$. This is a strong assumption for LSS (indeed, it does not hold for the DC microgrid case study considered in this thesis), where subsystems are often described as being weakly coupled and therefore their interconnection matrices being low rank [Lun92]. Furthermore, if we consider the subsystems to be structurally similar, defining the normal rank of the structural matrix $[A_{ij}] = n_{ij}$ equal for all $i, j \in \mathcal{N}$, would not solve Problem 3.2, since, *at best*, $n_{ij} = q_j$, thus not allowing for detection [Lemma A.1].

Thus, attacks that are stealthy to $\mathcal{O}_{ji}^{UIO}$ may not be isolatable by $\mathcal{O}_i^L$, depending on the structural characteristics of the interconnection matrices A_{ij} .

3.8 Plug-and-play design of $\mathcal{D}_i$

Having presented a distributed cyber-attack diagnoser $\mathcal{D}_i$ capable of detecting attacks injected on the communication lines linking the distributed controllers of LSSs, we now analyze its plug-and-play capabilities. Thus, we devise an algorithm to deal with events in which the set $\mathcal{N}$ of subsystems in an LSS changes.

Let us define $\mathcal{N}^+$ the set of subsystems of the LSS after an alteration: whenever $\mathcal{N} \subseteq \mathcal{N}^+$, a subsystem is said to have been “plugged-in” to the LSS; on the other hand if $\mathcal{N}^+ \subseteq \mathcal{N}$, a subsystem is said to have been “unplugged” from the system. The different plug-and-play operations are illustrated in Figure 3.3.

Remark 3.19. Note that the feasibility of the plug-in operation, critical in deciding whether a novel subsystem may disrupt the operation of the LSS, does not depend on the monitoring layer. Therefore, the detection architecture does not contribute to whether plugging-in a new subsystem is feasible or not [RBFTP16, BRFTP19]. Indeed, so long as the control law (3.6) is supposed to be designed such that plug-in and unplugging events can be dealt with automatically, the overall architecture can be said to have plug-and-play capabilities. There are a number of methods existing in literature for plug-and-play control design, as shown in Chapter 1.3.2. ◁

Whenever a subsystem $\mathcal{S}_j$ is unplugged from the LSS, all the diagnosers $\mathcal{D}_i$ of subsystems

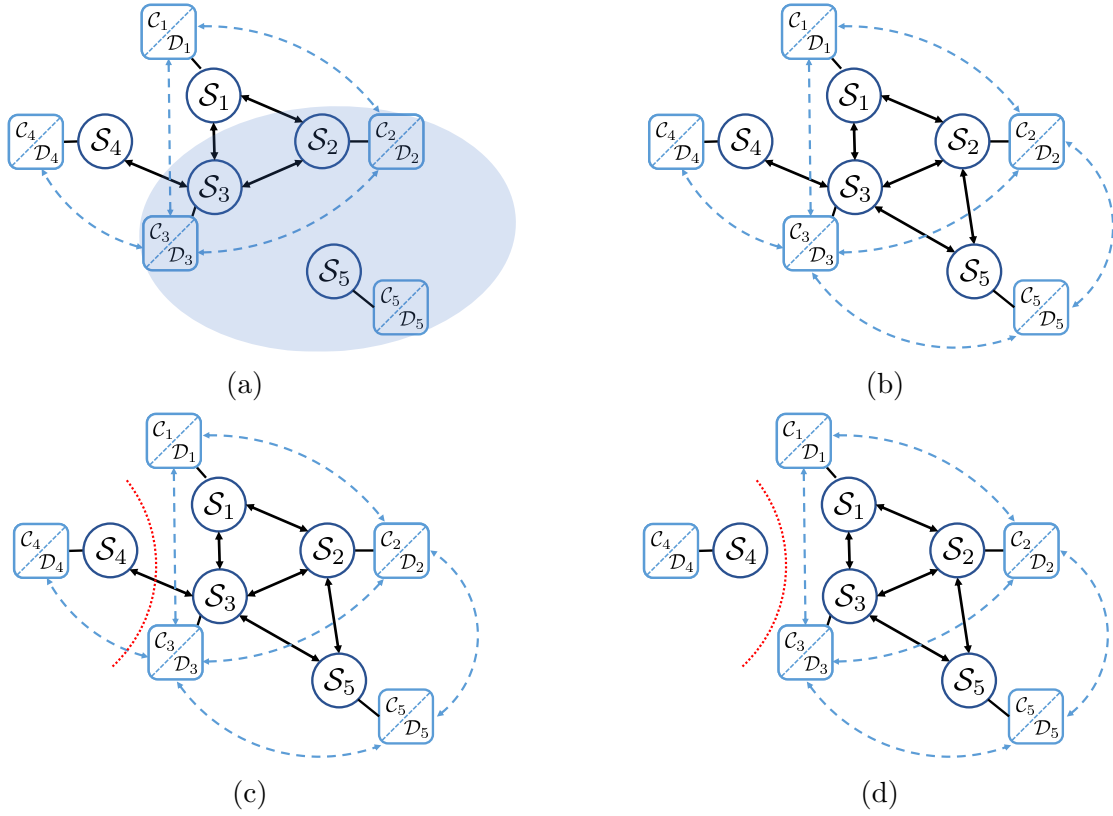


Figure 3.3: Different stages of plug-and-play operation of LSS as in Figure 3.1: (a) S_5 before plug-in – neighbours of S_5 , $\mathcal{N}_5 \doteq \{2, 3\}$, and subsystem to be plugged-in are highlighted; (b) LSS after plug-in of S_5 ; (c) LSS before unplugging S_4 ; (d) S_4 unplugged from LSS.

$\mathcal{S}_i, i \in \mathcal{N}_j$ must be updated such that their dependency on γ_{ji}^c is interrupted. To do this, the following operations are performed by $\mathcal{D}_i, \forall i \in \mathcal{N}_j$ following an unplugging event:

- a. remove j from $\mathcal{N}_i$;
- b. interrupt online operations of $\mathcal{O}_{ji}^{UIO}$;
- c. remove $A_{ij}C_j^\dagger \gamma_{ji}^c$ from the dynamics of $\hat{x}_i$ in $\mathcal{O}_i^L$.

On the other hand, whenever a subsystem $\mathcal{S}_j$ is added to the LSS, it must design its own $\mathcal{D}_j$, and all of its neighbours must update their diagnosers to account for $\mathcal{S}_j$. Specifically, before plug-in:

- The plugged-in subsystem $\mathcal{S}_j$ must:
 - a. receive model information (A_{ii}, E_i, C_i) as well as bounds $\bar{w}_i, \bar{v}_i$ from its neighbours $i \in \mathcal{N}_j$;

- b. share model information (A_{jj}, E_j, C_j) and bounds $\bar{w}_j, \bar{v}_j$ with $\mathcal{D}_i, i \in \mathcal{N}_j$;
 - c. design $\mathcal{O}_i^L$ and $\mathcal{O}_{ji}^{UIO}, \forall i \in \mathcal{N}_j$;
 - d. initialise $\bar{r}_j$ and $\bar{r}_{ij}$.
- All neighbouring subsystems $\mathcal{S}_i, i \in \mathcal{N}_j$ must:
- a. design module $\mathcal{O}_{ji}^{UIO}$ given received information (A_{jj}, E_j, C_j) ;
 - b. include A_{ij} in the dynamics of $\hat{x}_i$ within $\mathcal{O}_i^L$;
 - c. update computation of $\bar{r}_i$ and initialise $\bar{r}_{ji}$.

The above guarantee, once a subsystem is plugged-in or unplugged, that no attacks are falsely detected, and that the security of the LSS is preserved.

Remark 3.20. Note that, for the case of DC microgrids, the combination of the attack isolation methods delineated in Section 3.7 with the plug-and-play capabilities described here, presents an intuitive way of preserving some level of performance, which has been explored in the distributed fault-tolerant control literature (e.g. [RBFTP14, BCF⁺19]): indeed, once a specific communication link (j, i) is isolated as under attack, then $\mathcal{S}_i$ may remove $\mathcal{S}_j$ from its set of physical and communication neighbours. In this case, only $\mathcal{S}_i$ and $\mathcal{S}_j$ would need to reconfigure their controllers and diagnosers and current sharing is preserved, thanks to the plug-and-play nature of the consensus-based secondary controller presented in Chapter 2.1.2. Note that this is the case only so long as the connectivity of the graphs induced by the physical interconnections and communication coupling is maintained. ◁

3.9 Case study – DC microgrids

We now demonstrate the effectiveness of the architecture through simulation. Specifically, we consider a DC microgrid. Although having presented its modelling framework in Chapter 2, let us give a brief summary of its dynamics here, to address the design of modules $\mathcal{O}_{ji}^d$.

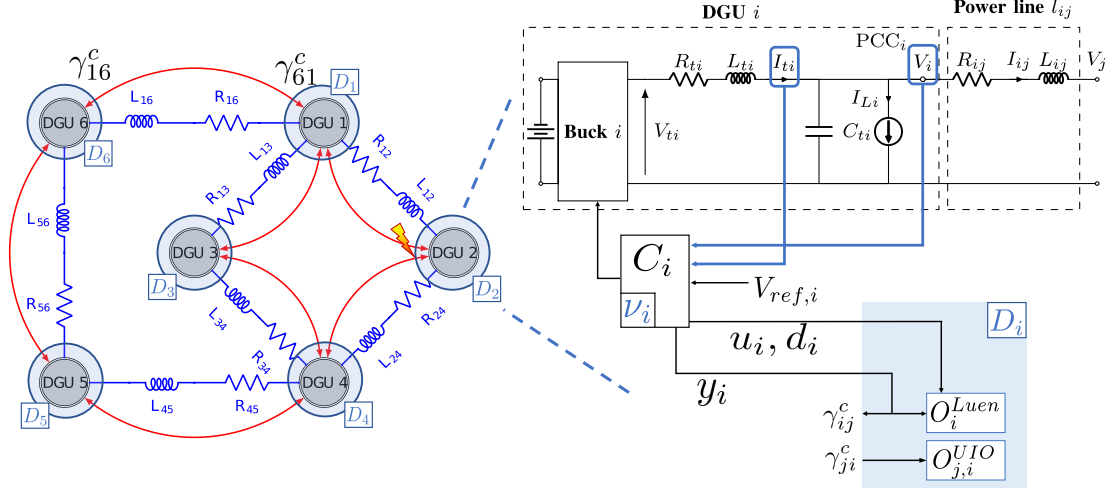


Figure 3.4: DC microgrid represented as a LSS with $N = 6$ DGUs as subsystems. [On the left] DC microgrid – each node represents a DGU, in blue the physical power lines between subsystems, in red the bi-directional communication links. [On the right] Representation of a distributed generation unit (composed of a power source, a Buck DC-DC converter, and an RLC filter), its controller unit $\mathcal{C}_i$ and the diagnostic unit $\mathcal{D}_i$.

3.9.1 DC microgrids – an overview

We consider a DC microgrid as being composed of N Distributed Generation Units (DGUs), each of which includes a voltage source, a DC-DC converter and a local load, as shown in Figure 3.4. Each DGU is coupled to its neighbours by RL power lines. There are two physical states, the voltage V_i at the point of common coupling (PCC), and the terminal current I_{ti} from the Buck converter.

To control the DC microgrid, we consider the distributed hierarchical plug-and-play controller defined in [TRFT18, TMGFT18], presented in Chapter 2.1.2. This controller is composed of a primary and a secondary controller, each achieving different objectives.

For voltage stabilisation, the primary layer of the regulator is constructed as a state feedback controller, introducing an internal integrator state ν_i to ensure that V_i tracks a reference $V_{ref,i}$. Therefore, we consider the overall state to be $x_i = [V_i, I_{ti}, \nu_i]^\top$. We suppose that the state of each DGU is fully measurable, and therefore that $p_i = n_i, \forall i \in \mathcal{N}$.

The secondary layer of the controller, on the other hand, is set to achieve current balancing.

To do this, the regulator computes a compensation term ΔV_i , used to define $V_{ref,i} \doteq V_{ref} + \Delta V_i$, with V_{ref} common to all DGUs. Specifically, ΔV_i is the result of the following consensus protocol

$$\Delta \dot{V}_i = k_I \sum_{j \in \mathcal{N}_i} a_{ij} \left(\frac{I_{tj}^c}{I_{tj}^s} - \frac{I_{ti}^m}{I_{ti}^s} \right) \quad (3.110)$$

where k_I is a scaling term common to all DGUs, I_{tj}^c is the state of DGU j as received by DGU i , and I_{ti}^m is the state measured by DGU i . Note that, by design, I_{ti}^m is subject to measurement noise, while $I_{tj}^c, j \in \mathcal{N}_i$, is susceptible to attacks. Finally, I_{ti}^s is an appropriately defined scaling factor, to account for different ratings for each DGU.

Exploiting the quasi stationary line approximation [see Assumption 2.2 in Chapter 2], the dynamics of each DGU can be written as the following LTI system

$$\begin{cases} \dot{x}_i = A_{ii}x_i + B_i u_i + \sum_{j \in \mathcal{N}_i} A_{ij}x_j + E_i d_i + w_i \\ y_i = C_i x_i + v_i \end{cases} \quad (3.111)$$

with $u_i = V_{ti}^{sw}$ the switching control input, $d_i = [I_{Li}, V_{ref,i}]^\top$ the exogenous inputs, and matrices:

$$A_{ii} \doteq \begin{bmatrix} -\sum_{j \in \mathcal{N}_i} \frac{1}{C_{ti} R_{ij}} & \frac{1}{C_{ti}} & 0 \\ -\frac{1}{L_{ti}} & -\frac{R_{ti}}{L_{ti}} & 0 \\ -1 & 0 & 0 \end{bmatrix}, \quad B_i \doteq \begin{bmatrix} 0 \\ \frac{1}{L_{ti}} \\ 0 \end{bmatrix}, \quad (3.112)$$

$$A_{ij} \doteq \begin{bmatrix} \frac{1}{C_{ti} R_{ti}} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad E_i \doteq \begin{bmatrix} -\frac{1}{C_{ti}} & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}. \quad (3.113)$$

Remark 3.21. As discussed in Chapter 2.1, the DC-DC converter in each DGU is a switched-mode power supply (SMPS), and therefore the dynamics of the DGU are not time invariant. However, through appropriate definition of input voltage V_{ti}^{sw} , the dynamics of the DGU can be seen as LTI. ◁

Recalling definition of $\bar{E}_i$ in (3.55), the unknown-input matrix is defined for each DGU as:

$$\bar{E}_i = \text{Im} \begin{bmatrix} B_i & E_i & A_{ij_1} & \cdots & A_{ij_{N_i}} \end{bmatrix}. \quad (3.114)$$

Given definition of the system matrices in (3.112), it is clear to see that

$$\text{rank } \bar{E}_i = \text{rank } C_i. \quad (3.115)$$

As shown in Appendix A.2.3, under this condition it is not possible to design a UIO capable of detecting attacks. Hence, to detect attacks on the communication between DGUs, it is necessary to augment the system dynamics as proposed in Appendix A.2.3. We here omit the definition of the augmented system to preserve clarity in the exposition. The details can be found in Appendix C.

3.9.2 Simulation setup

We considered a network of $N = 6$ distributed generation units, interconnected as shown in Figure 3.4. The simulations were carried out using the Specialized Power Systems Toolbox [MAT]. All DGUs were considered to have a nominal source voltage of $V_{si} = 60\text{V}$ and global voltage reference $V_{ref} = 48\text{V}$. All bidirectional Buck converters were considered as non-ideal IGBT switches, operating at a switching frequency $f_{sw} = 10\text{kHz}$, with snubbers¹⁵ suppressing large transients and protecting the equipment.

In simulation, the power lines interconnecting neighbouring DGUs were considered to be resistive and inductive, differently from Chapter 2.1, where the Quasi-Stationary Line approximation [Assumption 2.2] is used for modelling purposes. The values of the physical parameters of the DGUs and the power lines coupling neighbouring DGUs are given in Table 3.4.

Each DGU is equipped with a dynamical controller with a primary component used for voltage stabilization and balancing, and a secondary component for current sharing. The parameters of the control law are found in Table 3.5. Both electrical and controller parameters of the DC microgrid are taken from [TMGFT16].

¹⁵Snubbers are electronic devices used across a power switching device to suppress high voltage spikes.

Table 3.4: Electrical parameters of DC microgrid

Converter parameters				Power line parameters		
DGU	R_{ti} [Ω]	L_{ti} [mH]	C_{ti} [mF]	Connected DGUs (j, i)	R_{ij} [Ω]	L_{ij} [μ H]
$\mathcal{S}_1$	0.2	1.8	2.2	(1, 2)	0.05	2.1
$\mathcal{S}_2$	0.3	2.0	1.9	(1, 3)	0.07	1.8
$\mathcal{S}_3$	0.1	2.2	1.7	(1, 6)	0.1	2.5
$\mathcal{S}_4$	0.5	3.0	2.5	(2, 4)	0.04	2.3
$\mathcal{S}_5$	0.4	1.3	2.0	(3, 4)	0.06	1.0
$\mathcal{S}_6$	0.6	2.5	3.0	(5, 4)	0.08	1.8
				(5, 6)	0.08	3.0

Table 3.5: Parameters of distributed controllers

Primary control				Secondary control	
	$k_{i,1}$	$k_{i,2}$	$k_{i,3}$		
$\mathcal{C}_1$	−2.134	−0.163	13.553		
$\mathcal{C}_2$	−0.869	−0.050	48.285		
$\mathcal{C}_3$	−0.480	−0.108	30.673		
$\mathcal{C}_4$	−6.990	−0.175	102.960		
$\mathcal{C}_5$	−0.101	−0.010	16.393		
$\mathcal{C}_6$	−2.134	−0.163	13.553		

	$\mathcal{C}_i, \forall i \in \mathcal{N}$
I_{ti}^s	1
k_I	0.5

In the following, voltages are measured in [V], currents in [A]. The distributed controller exploits a virtual integrator state, measured in [V · s].

Process and measurement disturbances satisfy Assumption 3.1, with

$$\bar{w}_i = [0.05, 0.05, 0.01]^\top, \quad \bar{v}_i = [0.01, 0.01, 0]^\top, \quad \forall i \in \mathcal{N} \quad (3.116)$$

and are modelled as independent identically distributed random variables of uniform distribution between $[-\bar{s}, \bar{s}]$, for $s = \{v_{i,l}, w_{i,l}\}$, $l = \{1, 2, 3\}$.

Remark 3.22. The bound on the process disturbance $\bar{w}_{i,3}$ is chosen to be the same as the bound on the measurement disturbance $\bar{v}_{i,1}$. Indeed, although ν_i is an internal controller state,

its dynamics

$$\dot{v}_i = V_{ref,i} - V_i$$

are affected by the measurement error on the voltage sensor. $\triangleleft$

The gains of the Luenberger observers in $\mathcal{O}_i^L$ are computed such that

$$\sigma(A_{Li}) = \{-50, -100, -500\}$$

for all $i \in \mathcal{N}$. On the other hand, given definition of augmented $\bar{E}_j$ in Appendix C, the UIO matrices H_j are selected such that

$$S_j = I_3 \otimes \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$$

for all $j \in \mathcal{N}_i, i \in \mathcal{N}$. This ensures $S_j \bar{E}_j = 0$ for the augmented system. Matrices $\tilde{K}_j$ are then computed assigning the eigenvalues of $F_j = S_j A_{jj} - \tilde{K}_j C_j$:

$$\sigma(F_j) = \{-1, -1.5, -2, -2.5, -3, -3.5\}.$$

All other UIO matrices follow from their definition in (3.58).

In the following, we consider that the communication signals received by $\mathcal{S}_4$ are under attack. Specifically, two attack scenarios are discussed:

1. γ_{24}^c and γ_{34}^c are designed such that condition in Proposition 3.3 holds:
 - a. attacks satisfying Corollary 3.1 cooperatively;
 - b. both attacks designed to satisfy Corollary 3.2;
2. a covert attack is implemented on γ_{24}^c .

These specific attack scenarios are chosen to highlight the combined properties of $\mathcal{O}_i^L$ and $\mathcal{O}_{ji}^{UIO}$ in $\mathcal{D}_i$. Indeed, in Scenario 1. the attack signals are designed to be stealthy to $\mathcal{O}_i^L$, while the malicious agent in Scenario 2. injects attacks stealthy to $\mathcal{O}_{ji}^{UIO}$.

For both scenarios, at time $t = 0$ s all DGUs are initialised disconnected from all their neighbours, controlled only with the decentralised primary state feedback control law, tracking constant $V_{ref} = 48$ V. At $t = 2$ s, all DGUs are connected to each other through RL lines, the DC microgrid has topology in Figure 3.4, and the secondary controllers are activated to achieve current sharing. We consider that until this phase of the simulation all communication are healthy. Finally, at time $t = 8$ s, the attack is launched on the communication channels, as described above.

3.9.3 Scenario 1. – False data injection stealthy to $\mathcal{O}_i^L$

As pointed out in Section 3.2, and then repeated in Proposition 3.3, for attacks $\alpha_{ji}(t), j \in \hat{\mathcal{N}}(t)$ to be stealthy to $\mathcal{O}_i^L$, they must satisfy:

$$\sum_{j \in \hat{\mathcal{N}}_i(t)} A_{ij} C_j^\dagger \alpha_{ji}(t) = 0$$

As shown in Corollary 3.1 and Corollary 3.2, this condition can be achieved through coordination of attacks on multiple communication lines, or by exploiting structural rank deficiencies of the physical interconnection matrices A_{ij} . As anticipated, in the following we analyse both.

Scenario 1.a.: Stealthiness to $\mathcal{O}_i^L$ through coordination In this scenario, constant bias injection attacks are injected on communication lines γ_{24}^c and γ_{34}^c . The results are shown in Figure 3.5 and Figure 3.6.

Here, we randomly select the components of the attack vector α_{24} from a uniform distribution over the interval $[-0.02, +0.02]$. In order to satisfy

$$\sum_{j \in \hat{\mathcal{N}}_4(t)} A_{4j} C_j^\dagger \alpha_{j4}(t) = 0$$

the attack vector α_{34} is chosen in relation to α_{24} : Specifically, it is selected such that:

$$\alpha_{34,q} = -\frac{R_{34}}{R_{24}} \alpha_{24,q}$$

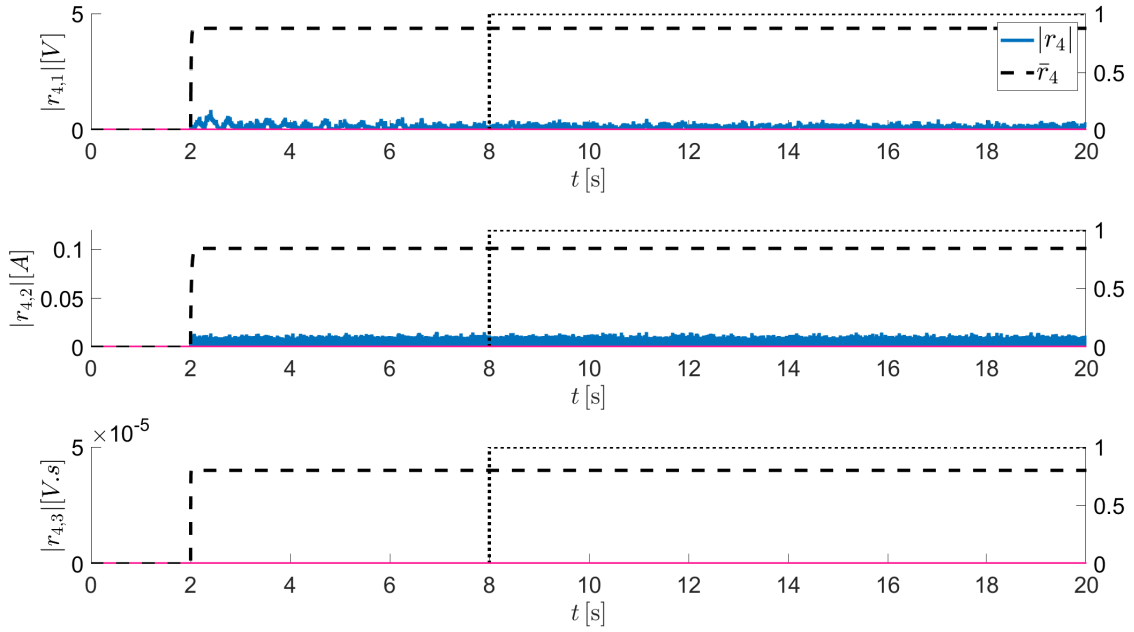


Figure 3.5: **[Scenario 1.a.: Coordinated attacks for stealthiness]**: Residuals and thresholds of $\mathcal{O}_4^L$ under stealthy false-data injection attack with attacks coordinating to achieve stealthiness. The false-data injection attack is indeed stealthy to $\mathcal{O}_4^L$. The residuals are shown in blue, while the thresholds are the dashed-black lines; furthermore, in dotted-black, a flag indicating the initial time of attack $T_a = 8$ s.

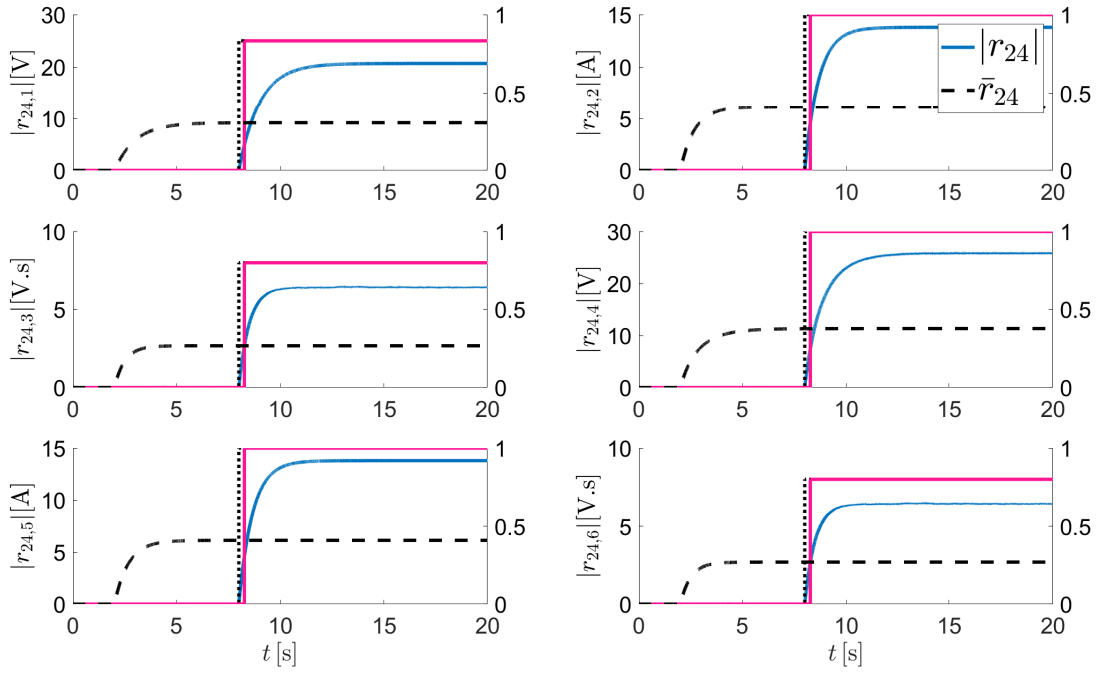
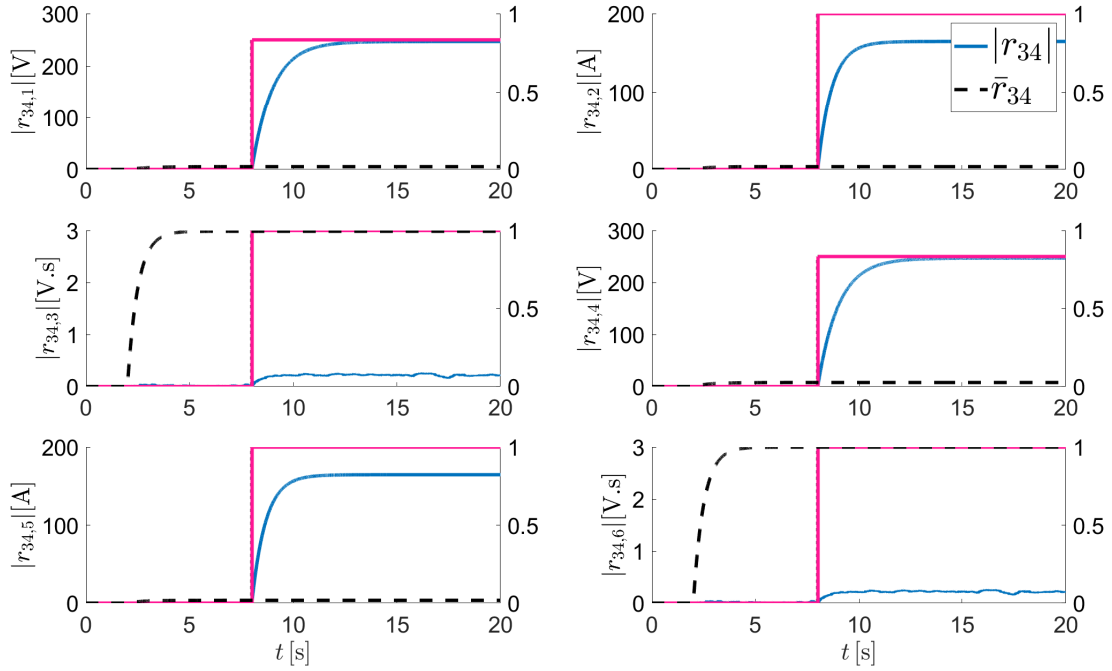
(a) Residuals and thresholds – $\mathcal{O}_{24}^{UIO}$ (b) Residuals and thresholds – $\mathcal{O}_{34}^{UIO}$

Figure 3.6: **[Scenario 1. a.: Coordinated attacks for stealthiness]**: Residuals and thresholds of $\mathcal{O}_{24}^{UIO}$ and $\mathcal{O}_{34}^{UIO}$ under stealthy false-data injection attack with attacks coordinating to achieve stealthiness. The colour scheme used is the same as that in Figure 3.5. Furthermore, a step function, in pink, indicates the initial time of the attack within the two modules, respectively.

where q is the component of γ_{ji}^c corresponding to the measurement of V_j – i.e. for the augmented system presented in Appendix C, $q = 4$. The remaining components of α_{34} are drawn from the uniform distribution within the interval above. The constant attack vectors are defined as:

$$\begin{aligned}\alpha_{24} &= [-0.0139, 0.0149, -0.0031, -0.0014, -0.0095, -0.0011]^\top, \\ \alpha_{34} &= [0.0037, 0.0185, 0.0174, 0.0021, 0.0178, 0.0180]^\top.\end{aligned}\tag{3.117}$$

In Figure 3.5 and Figure 3.6 we compare the residuals and thresholds of the local module $\mathcal{O}_4^L$ and the UIO modules $\mathcal{O}_{24}^{UIO}$ and $\mathcal{O}_{34}^{UIO}$. It is clear to see from Figure 3.5 that this coordinated attack strategy is stealthy to the Luenberger-observer-based module, while from Figure 3.6 it is evident that the two UIO-based modules monitoring γ_{24}^c and γ_{34}^c are capable of detecting the attacks. Specifically, with the initial time of attack set for both signals $T_a^{24} = T_a^{34} = 8$ s, the attack is detected by $\mathcal{O}_{24}^{UIO}$ and $\mathcal{O}_{34}^{UIO}$ at $T_d^{24} = 8.270$ s and $T_d^{34} = 8.015$ s, respectively.

Scenario 1.b.: Stealthiness to $\mathcal{O}_i^L$ through structural rank deficiencies Let us now validate through simulation the results presented in Corollary 3.2, shown in Figure 3.7 and Figure 3.8. Similarly to the previous scenario, we here introduce attacks on the communication lines γ_{ji}^c , for $i = 4$ and $j \in \{2, 3\}$. We consider the two attacks α_{24} and α_{34} to be non-zero from time $T_a^{24} = T_a^{34} = 8$ s, and we suppose that they alter the transmitted measurements of I_{t2}^c and I_{t3}^c , respectively. Specifically, we define the attack signals as

$$\begin{aligned}\alpha_{24} &= [0, 0, 0, 0, \tilde{\alpha}_{24}, 0]^\top \\ \alpha_{34} &= [0, 0, 0, 0, \tilde{\alpha}_{34}, 0]^\top\end{aligned}\tag{3.118}$$

where $\tilde{\alpha}_{24}$ and $\tilde{\alpha}_{34} \in \mathbb{R}$ are two arbitrary constants

$$\tilde{\alpha}_{24} = 0.04206, \quad \tilde{\alpha}_{34} = -0.02907.\tag{3.119}$$

Note that this attack is particularly harmful if undetected, as the received measurements $I_{tji}^c, j \in \mathcal{N}_i$ are used in the secondary controller of each DGU i to compute the reference compensation term ΔV_i , as shown in (3.110) and in Chapter 2. As $I_{tji}^c, j \in \mathcal{N}_i$ does not directly influence dynamics of $x_i, i \in \mathcal{N}$, these attacks satisfy the condition in Corollary 3.2.

In Figure 3.7, where we compare the residual r_4 to the threshold $\bar{r}_4$ as computed by $\mathcal{O}_4^L$, we see that the module exploiting only local information for its design is indeed unable to detect this attack. On the other hand, in Figure 3.8, we show how α_{24} and α_{34} are both detected promptly by $\mathcal{O}_{24}^{UIO}$ and $\mathcal{O}_{34}^{UIO}$ at time $T_d^{24} = 8.139$ s, and $T_d^{34} = 8.216$ s, respectively.

3.9.4 Scenario 2. – Covert attack against $\mathcal{O}_{24}^{UIO}$

In this scenario, we consider a single attacker injecting α_{24} onto communication between $\mathcal{C}_2$ and $\mathcal{C}_4$. Specifically, we consider a covert attack, with $\alpha_{24} = \Delta\gamma_{24}^a$, where $\Delta\gamma_{24}^a$ is the output of the system (3.95). The unknown input Δd_{24}^a is such that the Δx_{24}^a as observed by $\mathcal{O}_{24}^{UIO}$ have an altered operating point, with $\Delta I_{L24} = 2$ A.

In Figure 3.9, we compare the performance of the two modules $\mathcal{O}_4^L$ and $\mathcal{O}_{24}^{UIO}$. As expected, given that the attacker satisfies (3.95), the UIO-based module $\mathcal{O}_{24}^{UIO}$ does not detect the attack. On the other hand, Δx_{24}^a is such that it triggers detection by $\mathcal{O}_4^L$, as seen in Figure 3.9a, at time $T_d^4 = 8.001$ s.

3.10 Conclusions

In this chapter, we introduced a novel distributed monitoring architecture to endow control systems with cyber-attack detection capabilities. The proposed architecture shows how, exploiting different sets of trusted information within different detection modules, a local diagnosis module can be designed to detect cyber-attacks present in the communication network linking distributed controllers of a LSS.

We showed the properties of the two detection modules, as well as demonstrating their complementarity when combined. Furthermore, we present *structurally stealthy* attacks, i.e. those maliciously injected signals that can evade any detection strategy given a certain amount of information available.

Finally, through simulation results gathered from a realistic Simulink model of a DC microgrid, we show the efficacy of the proposed architecture.

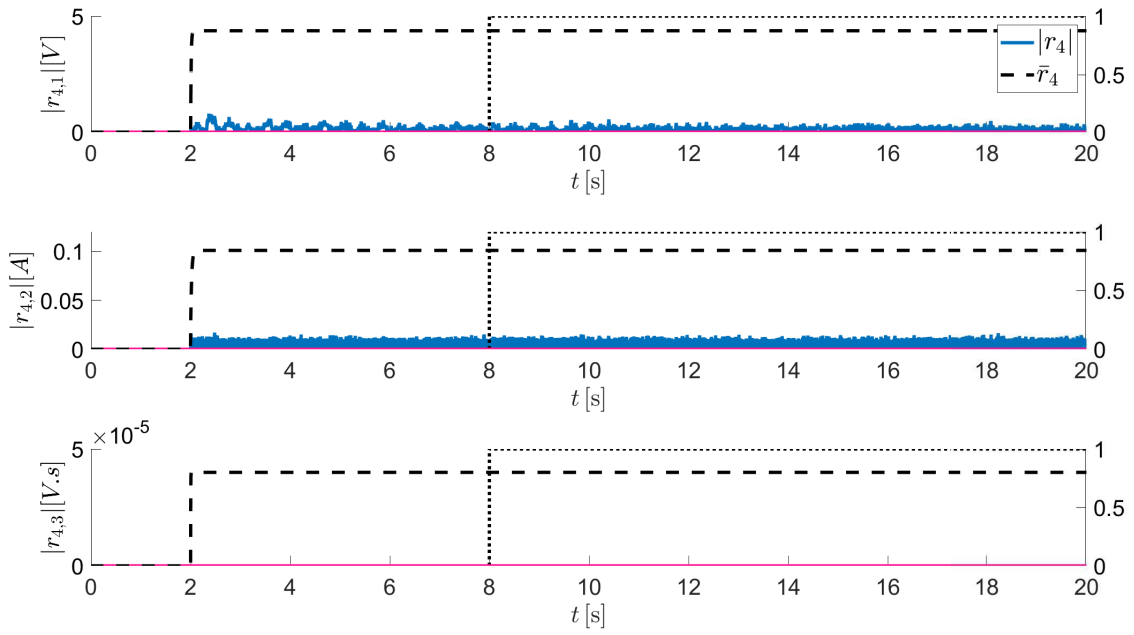


Figure 3.7: [**Scenario 1.b. Structural stealthiness**]: Residuals and thresholds of $\mathcal{O}_4^L$. The false-data injection attack α_{24} and α_{34} are designed to be stealthy to $\mathcal{O}_4^L$. The colour conventions are the same as in Figure 3.5.

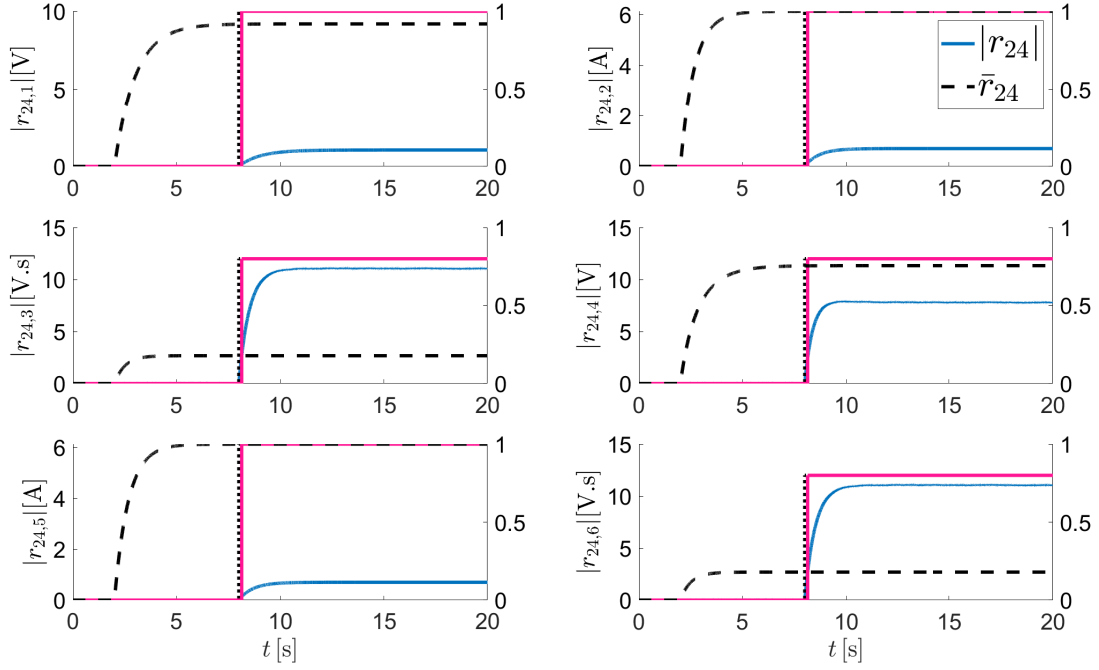
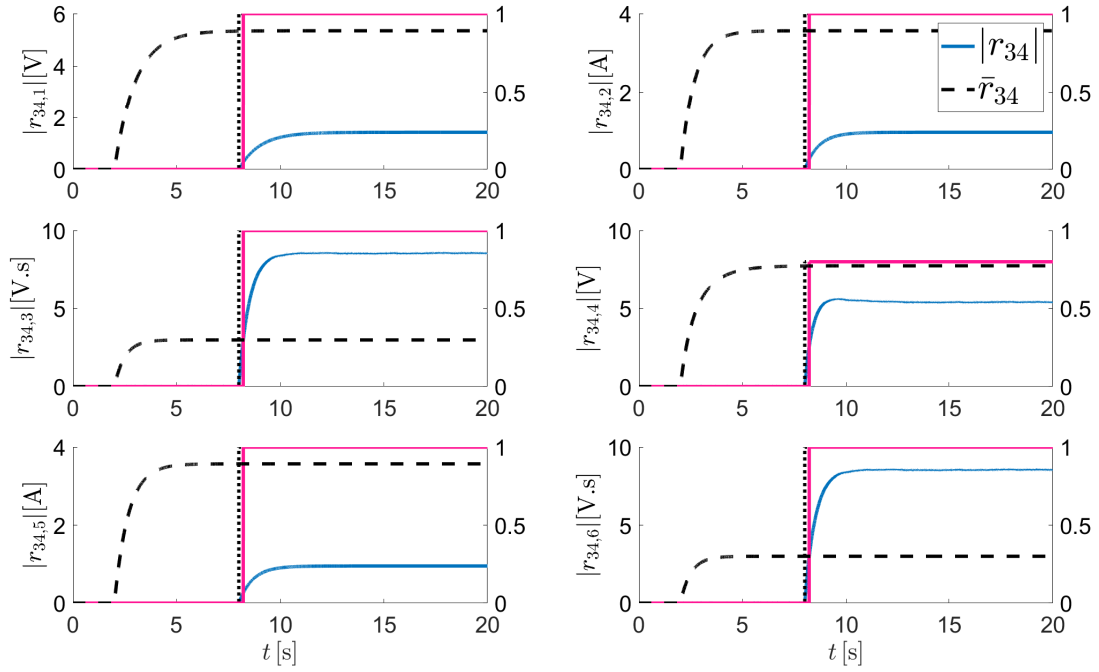
(a) Residuals and thresholds – $\mathcal{O}_{2,4}^{UIO}$ (b) Residuals and thresholds – $\mathcal{O}_{3,4}^{UIO}$

Figure 3.8: [Scenario 1.b. Structural stealthiness]: Residuals and thresholds of $\mathcal{O}_{24}^{UIO}$ and $\mathcal{O}_{34}^{UIO}$ under Scenario 1. The false-data injection attack is successfully detected by both modules. The figure shares the same colour conventions as Figure 3.6.

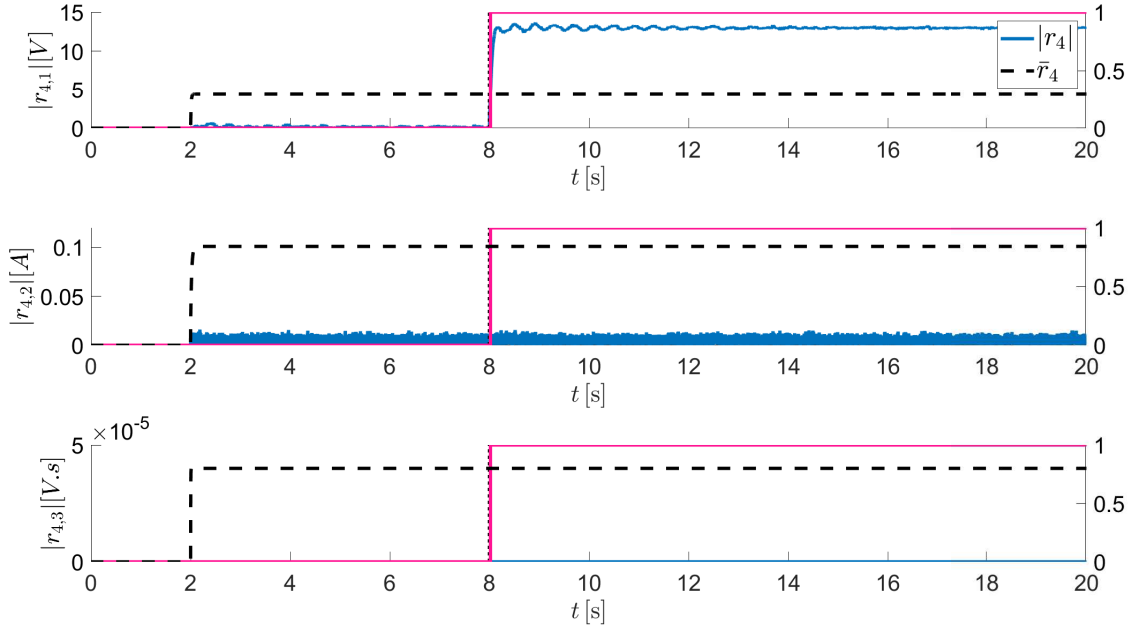
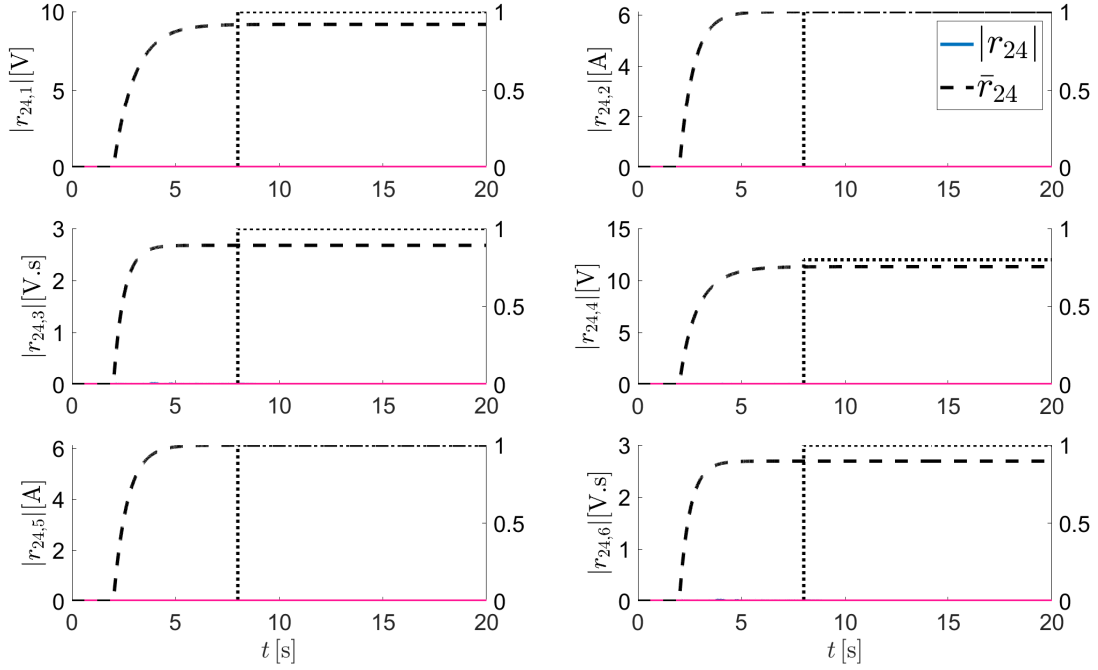
(a) Residuals and thresholds – $\mathcal{O}_4^L$ (b) Residuals and thresholds – $\mathcal{O}_{24}^{UIO}$

Figure 3.9: [Scenario 2 Covert attack against $\mathcal{O}_{24}^{UIO}$]: Residuals and thresholds of $\mathcal{O}_{24}^{UIO}$ and $\mathcal{O}_{34}^{UIO}$ under Scenario 2. The covert attack γ_{24}^c is successfully detected by $\mathcal{O}_4^L$ at time $T_d = 8.001$ s. The figure shares the same colour conventions as Figure 3.6.

Chapter 4

Distributed cyber-attack detection under stochastic uncertainties

In the previous chapter we have presented and analysed a distributed cyber-attack detection architecture for a LSS composed of N subsystems with continuous time dynamics

$$\begin{cases} \dot{x}_i = A_{ii}x_i + B_i u_i + E_i d_i + \sum_{j \in \mathcal{N}_i} A_{ij} x_j + w_i \\ y_i = C_i x_i + v_i. \end{cases} \quad (4.1)$$

Recall that each subsystem is considered to be subject to uncertainties $w_i(t)$ and $v_i(t)$ which are unknown-but-bounded. This implies that no knowledge of $w_i(t)$ or $v_i(t)$ is available *a priori*, apart from some positive bounds $\bar{w}_i$ and $\bar{v}_i$, satisfying

$$|w_i(t)| \leq \bar{w}_i, \quad |v_i(t)| \leq \bar{v}_i, \quad \forall t \geq 0. \quad (4.2)$$

Because of this, the properties of the distributed diagnosis architecture are computed based on a “worst case” scenario, i.e. on the known bounds $\bar{w}_i$ and $\bar{v}_i$. Although this guarantees the absence of false alarms, the properties of $\mathcal{D}_i$ as defined in Chapter 3 may be seen as conservative [BRFTP19]. In some applications, this may be useful, as interrupting operations for a false-alarm may be costly. In other contexts, however, it may be advantageous to allow for a less conservative detection test, in exchange for a non-zero probability of false alarm. Designing

such a distributed diagnoser for cyber-attack detection is the objective of this chapter.

To achieve this, therefore, we redefine the uncertainties w_i and v_i , providing them with some “structure”. Specifically, we characterise them as being realisations of a known stochastic distribution. The introduction of stochastic uncertainties in continuous time dynamical systems requires the use of a set of assumptions and tools which make analysis complex and algorithms cumbersome. We therefore break the continuity with respect to the problem formulation in Chapters 2 and 3, introducing an equivalent problem: distributed attack detection for large-scale interconnected systems modelled in discrete-time, subject to stochastic uncertainties. Thus, we define the dynamics of each subsystem as:

$$\begin{cases} x_i(k+1) = A_{ii}^d x_i(k) + B_i^d u_i(k) + E_i^d d_i(k) + \sum_{j \in \mathcal{N}_i^d} A_{ij}^d x_j(k) + w_i(k) \\ y_i(k) = C_i x_i(k) + v_i(k) \end{cases} \quad (4.3)$$

with disturbances $w_i(k)$ and $v_i(k)$ defined as discrete time realisations of white Gaussian noise, and $\mathcal{N}_i^d$ the set of subsystems neighbouring $\mathcal{S}_i$ given the discrete time dynamics. Matrices with superscript d represent the discrete time counterpart of those in (4.1). Furthermore, for any signal s_i , we abuse notation by defining $s_i(k) = s_i(kT_s)$, where $T_s \in \mathbb{R}_{>0}$ is the sampling period of the system. The full problem definition is given in Section 4.1.

To maintain a connection to the problem formulation in Chapter 3, we consider (4.3) to be the discretisation of (4.1). Specifically, we assume:

Assumption 4.1. *The sampling of all subsystems $\mathcal{S}_i$ is synchronous.* $\triangleleft$

Assumption 4.2. *We suppose a sampling-and-hold approach to discretisation, i.e. the following holds:*

$$u_i(t) = u_i(kT_s), \quad d_i(t) = d_i(kT_s), \quad x_j(t) = x_j(kT_s), \quad \forall j \in \mathcal{N}_i. \quad (4.4)$$

for all $t \in [kT_s, (k+1)T_s)$, for all $k \in \mathbb{N}_0$. $\triangleleft$

These assumptions allow us to compute the discretisation of the LSS dynamics subsystem-by-subsystem, thus preserving the sparsity of the interconnections between subsystems unchanged, i.e. for all subsystems $\mathcal{S}_i$ the discrete time set of neighbours satisfies $\mathcal{N}_i^d \equiv \mathcal{N}_i$.

Note furthermore that discretisation of the subsystem dynamics does not only allow us to introduce stochastic uncertainties in the dynamics of each subsystem, but is also consistent with the implementation of digital control and communication resources in large-scale systems.

The work presented in this chapter is based on the contribution:

- [BGBP19] A. Barboni, A. J. Gallo, F. Boem, T. Parisini, “A distributed approach for the detection of covert attacks in interconnected systems with stochastic uncertainties”, 58th Annual Conference on Decision and Control (CDC), pp. 5623–5628, 2019

which however presented a different problem formulation: indeed, differently from what is considered in this chapter, the objective in [BGBP19] was to design a distributed detection architecture capable of detecting local covert attacks. The class of *local covert attacks*, introduced by [BRBP19], is equivalent to the class of covert attacks in Chapter 2 in the context of large-scale systems; however, local covert attacks suppose that an attacker with knowledge of the dynamics of $\mathcal{S}_i$ performs a covert attack only against that subsystem, i.e. changing u_i and y_i . Apart from a difference in attack strategy, in [BGBP19] the communication between controllers $\mathcal{C}_i$ was considered to be secure.

Moreover, it is worth noting that a problem similar to the one analysed in [BGBP19] was presented preliminarily in [AKP18], then extended in [AKP20]. In this work, once more, the authors analyse attacks on local subsystem control inputs u_i , without considering the possibility of attacks on the communication network between controllers.

As mentioned, the objective of this chapter is to present a distributed cyber-attack detection scheme capable of detecting attacks on the communication between distributed controllers in the presence of stochastic uncertainties. We here focus on the properties of the diagnoser which depend on the stochastic nature of the formulation of the problem, without focusing our attention on the overall structure of the diagnosis architecture, as it closely resembles that presented in Chapter 3. Indeed, we introduce a distributed diagnoser $\mathcal{D}_i$ composed of two modules $\mathcal{O}_i^d$ and $\mathcal{O}_{ji}^d$ that, similarly to $\mathcal{O}_i^L$ and $\mathcal{O}_{ji}^{UIO}$ defined in Chapters 3.4 and 3.5, respectively, rely on an estimate of the local state x_i and of the neighbouring states $x_j, j \in \mathcal{N}_i$

to perform detection. Hence, the classes of attacks each module can detect are similar to the diagnosis modules offered in Chapter 3.

The primary focus of this chapter is therefore the analysis of the properties of the estimation scheme used to compute $\hat{x}_i$ estimating x_i in $\mathcal{O}_i^d$. It is well established in literature that distributed estimators of large-scale interconnected systems subject to stochastic uncertainties (often called partitioned-based), such as the partition-based Kalman filter presented in [CCSZ08, FC16] or the estimator in [BRFTP19], require the introduction of approximating methods to compute the estimation error covariance, and therefore the feedback gain. Specifically, these distributed estimators require some level of approximation to compute the statistical properties of the estimation error $\epsilon_i = x_i - \hat{x}_i$ in a distributed fashion. Indeed, it is shown that if the graph induced by the dynamics of the large-scale system is connected, the variance of ϵ_i depends on those of all other¹ estimation errors $\epsilon_k, k \in \mathcal{N}$.

It is worth noting explicitly that the estimator in [BRFTP19] is used within the context of a distributed plug-and-play architecture for fault detection and isolation for large-scale systems in the presence of stochastic uncertainties.

To avoid this, we design the module $\mathcal{O}_i^d$ by adapting the unbiased minimum-variance filter² proposed in [GDM07], considering the physical coupling between subsystems to be an unknown disturbance. Thus, our design is fully decentralised, and we remove the dependency of ϵ_i on $\epsilon_j, j \in \mathcal{N}_i$. On the other hand, $\mathcal{O}_{ji}^d$ is designed adapting the same unbiased filter, defined for a different set of unknown inputs.

4.1 Problem formulation

4.1.1 Large-scale interconnected subsystems

Let us start by formally defining the problem addressed in this chapter. To improve legibility, we drop the use of the superscript d used in (4.3) to differentiate the discrete-time dynamics

¹Note that if the graph induced by the physical coupling between subsystems is not undirected, the variance of ϵ_i is influenced by all those subsystems which are present in the same strongly-coupled subgraph [Bul18].

²We note that the problem of estimating the state of a system subject to unknown inputs and stochastic uncertainties has drawn significant interest in literature, with fundamental contributions including [Kit87, HP98b], as well as the aforementioned [GDM07]. An in depth description of the filter presented in [GDM07] can be found in Appendix A.3.

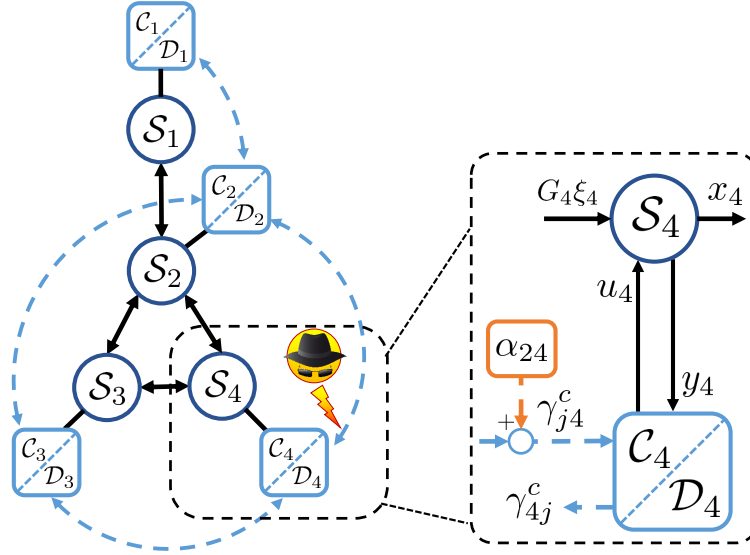


Figure 4.1: Representation of a large-scale system as a network of $N = 4$ subsystems with an attack on the communication received by $\mathcal{S}_4$ from $\mathcal{S}_2$. [On the left] Networked subsystems. Each subsystem is equipped with a controller $\mathcal{C}_i$ and a diagnoser $\mathcal{D}_i$. The physical coupling between subsystems is represented by the black arrows, the (secure) link between plants and controllers are the black lines, while the communication network amongst controllers is shown through the dashed-blue arrows. [On the right] Subsystem $\mathcal{S}_4$ with its control and monitoring layer and subject to attack signal α_{24} . The same colour convention as the left is used.

from the continuous-time dynamics in (4.1).

We consider a large-scale system such as the one represented in Figure 4.1. We suppose the LSS is divided into N non-overlapping and interconnected subsystems $\mathcal{S}_i, i \in \mathcal{N} \doteq \{1, \dots, N\}$. Each subsystem is taken to be physically coupled with a set of N_i subsystems $\mathcal{S}_j, j \in \mathcal{N}_i \doteq \{j \in \mathcal{N} | \partial x_i / \partial x_j \neq 0\} \subseteq \mathcal{N}$, and is described by the dynamics:

$$\begin{cases} x_i(k+1) = A_{ii}x_i(k) + B_i u_i(k) + \sum_{j \in \mathcal{N}_i} A_{ij}x_j(k) + E_i d_i(k) + w_i(k) \\ y_i(k) = C_i x_i(k) + v_i(k) \end{cases} \quad (4.5)$$

where $x_i \in \mathbb{R}^{n_i}$ is the state of $\mathcal{S}_i$, $y_i \in \mathbb{R}^{p_i}$ its output, and $u_i \in \mathbb{R}^{m_i}$ its control input. The variable $d_i \in \mathbb{R}^{q_i}$ represents the known exogenous inputs acting upon each subsystem. All matrices in (4.5) are supposed known and of appropriate dimensions. The signals w_i and v_i model the process and measurement noise, respectively, with the following characteristics:

Assumption 4.3. *Process and measurement noise signals w_i and v_i are mutually uncorrelated*

and independent identically distributed white Gaussian processes

$$w_i(k) \sim \mathcal{N}(0, W_i), \quad v_i(k) \sim \mathcal{N}(0, V_i), \quad \forall k \geq 0, \quad (4.6)$$

with $W_i \succcurlyeq 0$ and $V_i \succcurlyeq 0$ two known covariance matrices. The initial state $x_i(0) \doteq x_i^0$ is a realization of the normal distribution

$$x_i^0 \sim \mathcal{N}(\bar{x}_i^0, \Pi_i^0), \quad (4.7)$$

and is independent of $w_i(k)$ and $v_i(k)$ for all k , with mean $\bar{x}_i^0$ and covariance $\Pi_i^0 \succcurlyeq 0$ known. $\triangleleft$

Remark 4.1. Here we use $\mathcal{N}(s, S)$ as an operator to define a multivariable Gaussian distribution with mean $s \in \mathbb{R}^{n_s}$ and covariance matrix $S \in \mathbb{R}^{n_s \times n_s}$, without risk of confusion with $\mathcal{N}$, the set of all indices of subsystems composing the LSS. $\triangleleft$

Assumption 4.4. For all subsystems $\mathcal{S}_i$, $i \in \mathcal{N}$, (A_{ii}, C_i) is observable, and C_i is full row rank, i.e. $\text{rank } C_i = p_i \leq n_i$. $\triangleleft$

Assumption 4.5. For each subsystem $\mathcal{S}_i$, $j \in \mathcal{N}_i$ implies that $i \in \mathcal{N}_j$. $\triangleleft$

4.1.2 Distributed control architecture and cyber attack detection

The control input u_i is considered to be computed by a distributed controller $\mathcal{C}_i$ as:

$$u_i = \kappa_i(y_i, \gamma_{ji, j \in \mathcal{N}_i}^c), \quad (4.8)$$

where $\kappa_i(\cdot, \cdot)$ is some operator representing the control law, making use of the local output signal y_i as well as the communicated information $\gamma_{ji, j \in \mathcal{N}_i}^c$. To implement this distributed control architecture, a communication network is considered to be integrated in the control architecture, such that neighbouring controllers may transmit and receive information vectors γ_{ji}^c , which $\mathcal{C}_i$ receives from all its neighbours $j \in \mathcal{N}_i$.

Assumption 4.6. The graph induced by the communication network defined by $\mathcal{G}^c \doteq \{\mathcal{N}, \mathcal{E}^c\}$

is such that the set of edges $\mathcal{E}^c \equiv \mathcal{E}^p$, where $\mathcal{E}^p$ is the set of edges induced by the physical coupling between subsystems $\mathcal{G}^p \doteq \{\mathcal{N}, \mathcal{E}^p\}$, itself defined as $\mathcal{E}^p \doteq \{(j, i) | j \in \mathcal{N}_i, i \in \mathcal{N}\}$. $\triangleleft$

Remark 4.2. Note that, similarly to Assumption 3.4, we consider the topology of the graph induced by the communication network to be the same as that of the graph induced by the physical coupling among subsystems. This allows the analysis and notation to be simplified. Similarly to what we have shown in Chapter 3.3.4, this assumption is not necessary, although it is necessary that $\mathcal{E}^p \subseteq \mathcal{E}^c$ for the architecture to function properly. $\triangleleft$

The exchanged information between subsystems is defined as:

$$\gamma_{ji}^c(k) = \hat{x}_j(k) + \beta_{ji}(k - k_a^{ji})\alpha_{ji}(k) \quad (4.9)$$

where $\hat{x}_j$ is the estimate³ of the state x_j computed by $\mathcal{D}_j$, and $k_a^{ji} > 0$ the initial instant of attack. Function $\beta_{ji} : \mathbb{R} \rightarrow \mathbb{R}$ is, as in previous chapters, an attacker-defined modulation function, here defined as:

$$\beta_{ji}(k) = \begin{cases} (1 - b_{ji}^k) & k \geq 0 \\ 0 & k < 0 \end{cases} \quad (4.10)$$

with $b_{ji} \in (0, 1]$; the closer parameter b_{ji} is to $b_{ji} \approx 0$, the more abrupt the attack profile. The attack function $\alpha_{ji}(\cdot) : \mathbb{N}_0 \rightarrow \mathbb{R}^{n_j}$ is defined by the attacker and is presumed to be deterministic. Finally, let us define the initial time of attack

$$k_a^i \doteq \min_{j \in \mathcal{N}_i} \{k_a^{ji}\}. \quad (4.11)$$

Thus, the system is said to behave *nominally*, or to be *healthy*, for all $k \in \mathcal{K}_i^{\text{nom}} \doteq \{0, \dots, k_a^i - 1\}$, and to be under attack for all other time instances.

Remark 4.3. Note, similarly to Assumption 3.7 in Chapter 3, we consider that the signals $u_i(k)$ and $y_i(k)$ cannot be maliciously altered by an attacker. Hence, they are considered *trusted*, and thus exploited for cyber-attack detection in $\mathcal{D}_i$. This, as mentioned in Remark 3.5, is consistent

³Details on the design of $\hat{x}_i$ in $\mathcal{D}_i$ are given in Section 4.2.

with the systems we consider throughout this thesis, i.e. power-electronics-enabled distribution networks, such as DC microgrids. $\triangleleft$

We can now introduce the problem of distributed⁴ cyber-attack detection for interconnected systems with stochastic uncertainties.

Problem 4.1. Consider $\mathcal{S}_i$ a subsystem of an LSS, with dynamics as in (4.5) and subject to stochastic uncertainties w_i and v_i with properties defined in Assumption 4.3. Suppose each $\mathcal{S}_i$ is controlled by distributed controllers $\mathcal{C}_i$ with control protocol as in (4.8). Take the communication signal γ_{ji}^c transmitted between neighbouring controllers to be defined as in (4.9), and thus to be subject to attacks. Design a distributed detection architecture, equipping each subsystem $\mathcal{S}_i$ with a diagnoser $\mathcal{D}_i$ capable of detecting an attack $\alpha_{ji}(k) \neq 0$, $j \in \mathcal{N}_i$. $\triangleleft$

In the following we present the modules composing the distributed diagnoser $\mathcal{D}_i$, namely $\mathcal{O}_i^d$ and $\mathcal{O}_{ji}^d$, exploiting the distributed architecture developed in Chapter 3.

Similarly to $\mathcal{O}_i^L$ and $\mathcal{O}_{ji}^{UIO}$ presented previously, $\mathcal{O}_i^d$, on the one hand, uses trusted output measurements y_i to estimate the local state x_i , thus linking detection to the physical coupling between subsystems. On the other hand, $\mathcal{O}_{ji}^d$ estimates the state of the neighbouring subsystems to validate the trajectory of γ_{ji}^c . This architecture is illustrated in Figure 4.2.

4.2 Decentralized state and coupling estimation

We propose the design of a decentralized estimation filter which can estimate x_i independently of all neighbour information to be included in $\mathcal{O}_i^d$. Specifically, by making the estimate $\hat{x}_i$ independent of all information γ_{ji}^c received from $\mathcal{S}_j, j \in \mathcal{N}_i$, the statistical properties of the estimation error $\epsilon_i \doteq x_i - \hat{x}_i$ depend only on characteristics local to $\mathcal{S}_i$. Indeed, the decentralized estimation error depends only on the mean and variance of $w_i(k)$, $v_i(k)$ and x_i^0 , defined in Assumption 4.3. Although this allows for local computation of error variance, it is obvious that by making the estimate $\hat{x}_i$ independent of the received information γ_{ji}^c implies that it is also independent of attacks α_{ji} , and thus not suitable for detection.

⁴Recall the definition of distributed module in Definition 3.1.

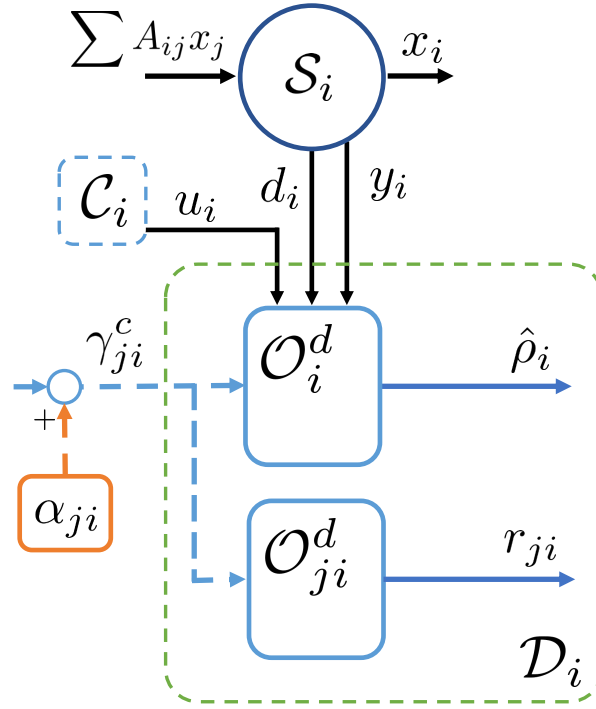


Figure 4.2: Representation of the diagnoser $\mathcal{D}_i$ designed to detect cyber-attacks, thus solving Problem 4.1, including both $\mathcal{O}_i^d$ and $\mathcal{O}_{ji}^d$.

To get around this, we introduce two different estimates $\hat{\xi}_i$ and $\check{\xi}_i$ of the aggregate physical coupling ξ_i between $\mathcal{S}_i$ and its neighbours $\mathcal{S}_j, j \in \mathcal{N}_i$, which we can rewrite as in (4.12).

Of the two estimates of ξ_i , $\hat{\xi}_i$ is obtained directly from the design of the decentralized filter estimating x_i , as shown in (4.14): it is therefore independent of all neighbouring subsystems, and thus of α_{ji} . On the other hand, estimate $\check{\xi}_i$ is computed from the received measurements γ_{ji}^c as in (4.32), and hence it is biased by the presence of nonzero attacks α_{ji} on the communication between controllers. Therefore, a residual-like signal $\hat{\rho}_i = \hat{\xi}_i - \check{\xi}_i$ is introduced and exploited to detect cyber-attacks.

Given that the local estimation filter estimating x_i is designed to be decentralized, we treat the physical coupling term in (4.5) as an unknown input. Let us introduce ξ_i as a variable modelling the aggregate interconnection affecting dynamics (4.5):

$$G_i \xi_i \doteq \sum_{j \in \mathcal{N}_i} A_{ij} x_j, \quad (4.12)$$

with $\xi_i \in \mathbb{R}^{g_i}$ the unknown input vector, and the matrix G_i defined as a basis for $\text{Im} \left[\text{row}_{j \in \mathcal{N}_i}(A_{ij}) \right]$.

Note that given its definition G_i is full-column-rank. We introduce the following assumptions on the structure of G_i in relation to the output matrix C_i .

Assumption 4.7. *Output matrices C_i are such that*

$$\text{rank}(C_i G_i) = \text{rank}(C_i) \quad (4.13)$$

for all subsystems $\mathcal{S}_i$, $i \in \mathcal{N}$. ◁

Remark 4.4. Note that Assumption 4.7 is required for decoupled state estimation in the presence of unknown inputs both in the stochastic and deterministic cases (indeed, recall Assumption 3.10). It is in fact used to ensure input observability of ξ_i .

Furthermore, the redefinition of the physical coupling as $G_i \xi_i$ is necessary, as indeed only

$$g_i \doteq \text{rank} \left[\text{row}_{j \in \mathcal{N}_i} (A_{ij}) \right]$$

components of the dynamic interconnection can be reconstructed from y_i . ◁

To obtain a decentralized estimate of x_i and ξ_i , we modify the minimum-variance unbiased filter defined in [GDM07]. Let us define $\hat{x}_i$ and $\hat{\xi}_i$ as the estimates of the local state x_i and the aggregate physical coupling ξ_i , respectively. These are computed following the dynamics:

$$\hat{x}_i(k|k-1) = A_{ii}\hat{x}_i(k-1|k-1) + B_i u_i(k-1) + E_i d_i(k-1) \quad (4.14a)$$

$$\hat{\xi}_i(k-1|k) = M_i(k) (y_i(k) - C_i \hat{x}_i(k|k-1)) \quad (4.14b)$$

$$\hat{x}_i^*(k|k) = \hat{x}_i(k|k-1) + G_i \hat{\xi}_i(k-1|k) \quad (4.14c)$$

$$\hat{x}_i(k|k) = \hat{x}_i^*(k|k) + K_i(k) (y_i(k) - C_i \hat{x}_i^*(k|k)) \quad (4.14d)$$

where matrices $M_i(k)$ and $K_i(k)$, defined in (4.15) and (4.16), guarantee that $\hat{x}_i(k|k)$ and $\hat{\xi}_i(k-1|k)$ are minimum-variance unbiased (MVU) estimates of $x_i(k)$ and $\xi_i(k)$.

The filter in (4.14) estimates the local state and unknown input in four stages: initially, in (4.14a), it computes the best-case prediction of the state $x_i(k)$ from information available at time $k-1$, and without any knowledge of the unknown input $\xi_i(k-1)$; following this,

once measurement $y_i(k)$ is available, an estimate $\hat{\xi}_i(k-1|k)$ can be computed, by using the innovation obtained by comparing $y_i(k)$ to the expected output $C_i\hat{x}_i(k|k-1)$, as in (4.14b); thus, by including the estimate of the unknown input, the prediction is updated, computing $\hat{x}_i^*(k|k)$ in (4.14c); finally, by exploiting the innovation $y_i - C_i\hat{x}_i^*$ and the appropriately defined gain $K_i(k)$, the decentralized MVU estimate of the local state is computed in (4.14d).

The matrices $M_i(k)$ and $K_i(k)$ in (4.14) are designed to minimize the variance of the estimates of $\hat{x}_i$ and $\hat{\xi}_i$ and are defined by:

$$M_i(k) \doteq (F_i^\top H_i(k)^{-1} F_i)^{-1} F_i^\top H_i(k)^{-1} \quad (4.15)$$

$$K_i(k) \doteq (A_{ii}\Pi_i(k-1)A_{ii}^\top + W_i)C_i^\top H_i(k)^{-1}, \quad (4.16)$$

where $\Pi_i(k) \in \mathbb{R}^{n_i \times n_i}$ is the variance matrix of the local estimation error $\epsilon_i(k) \doteq x_i(k) - \hat{x}_i(k)$, given in the following, the matrix H_i is defined as

$$H_i(k) = C_i (A_{ii}\Pi_i(k-1)A_{ii}^\top + W_i) C_i^\top + V_i, \quad (4.17)$$

and $F_i = C_i G_i$. Further details on the minimum-variance unbiased estimate proposed in [GDM07] are provided in Appendix A.

Remark 4.5. In (4.14), and throughout the remainder of this chapter, we include time indices to define explicitly when it is possible to compute an estimate of either $x_i(k)$ or $\xi_i(k)$. Specifically, note that (4.14b) depends on delayed information, as information on $\xi_i(k-1)$ is only available from measurements $y_i(k)$. This is to be expected, as the dynamical relationship between ξ_i and x_i in (4.5) implies that $x_i(k)$ is affected by $\xi_i(\kappa)$, $\kappa \in \{0, \dots, k-1\}$. Furthermore, it is consistent with results on unknown input reconstruction, e.g. [CT98, BFP09]. While this introduces a single step time delay in the computation of estimate $\hat{\xi}_i$, it does not introduce a delay in the estimate of x_i , as all information necessary for the computation of $\hat{x}_i(k)$ is available by time k . $\triangleleft$

Following the considerations in Remark 4.5, to simplify notation we rewrite $\hat{x}_i(k) = \hat{x}_i(k|k)$,

and rewrite the four steps defined in (4.14) in compact form:

$$\hat{x}_i(k) = \bar{A}_i(k) [A_i \hat{x}_i(k-1) + B_i u_i(k-1)] + \bar{L}_i y_i(k) \quad (4.18)$$

$$\hat{\xi}_i(k-1|k) = M_i(k) [y_i(k) - C_i (A_i \hat{x}_i(k-1) + B_i u_i(k-1))] \quad (4.19)$$

where

$$\bar{A}_i(k) \doteq (I - K_i(k)C_i)(I - G_i M_i(k)C_i) \quad (4.20a)$$

$$\bar{L}_i(k) \doteq K_i(k) + (I - K_i(k)C_i)G_i M_i(k). \quad (4.20b)$$

Note that the two matrices in (4.20) are such that:

$$\bar{A}_i(k) = I_{n_i} - \bar{L}_i(k)C_i. \quad (4.21)$$

Let us now introduce the main result in [GDM07], where the authors prove that the estimates $\hat{x}_i(k)$ and $\hat{\xi}_i(k-1|k)$ in (4.18) with filtering matrices $M_i(k)$ and $K_i(k)$ defined in (4.15)–(4.16), are minimum variance unbiased estimates of $x_i(k)$ and $\xi_i(k-1)$, respectively.

Theorem 4.1 (Theorem 12, [GDM07]). Consider the joint input and state estimator in (4.14), where $M_i(k)$ satisfies:

$$M_i(k)C_i G_i = I_{g_i}, \quad \forall k \geq 0. \quad (4.22)$$

If $M_i(k)$ and $K_i(k)$ are designed as in (4.15)–(4.16), then $\hat{x}_i$ and $\hat{\xi}_i$ are unbiased estimates of x_i and ξ_i , minimizing the mean square error over the class of linear unbiased estimates, based on $\bar{x}_i^0$ and $y_i(s), 0 \leq s \leq k$. $\square$

4.2.1 Local state estimate

Let us start with analysis of the state estimation error and residual, $\epsilon_i \doteq x_i - \hat{x}_i$ and $r_i \doteq y_i - C_i \hat{x}_i$, in nominal conditions, i.e. for all $k \in \mathcal{K}_i^{\text{nom}} \doteq \{1, \dots, k_a^i\}$. Given the definition of

dynamics of $\hat{x}_i$ in compact form in (4.18), the estimation error dynamics are:

$$\begin{cases} \epsilon_i(k) = \bar{A}_i(k) [A_{ii}\epsilon_i(k-1) + G_i\xi_i(k-1) + w_i(k-1)] - \bar{L}_i(k)v_i(k) \\ \quad = \bar{A}_i(k) [A_{ii}\epsilon_i(k-1) + w_i(k-1)] - \bar{L}_i(k)v_i(k) \\ r_i(k) = C_i\epsilon_i(k) + v_i(k). \end{cases} \quad (4.23)$$

Note that $\epsilon_i(k)$ is not influenced by $G_i\xi_i(k-1)$. Indeed, given that definition of $M_i(k)$ in (4.15) satisfies condition (4.22) in Theorem 4.1, we can simplify

$$\bar{A}_i(k)G_i = (I - K_i(k)C_i)(I - G_iM_i(k)C_i)G_i = 0, \quad \forall k \geq 0,$$

in turn leading to (4.23). Indeed, the error must not be influenced by ξ_i for $\hat{x}_i$ to be an unbiased decentralized estimate of x_i .

Given Assumption 4.3, the estimation error and the residual can be seen to be random variables with Gaussian distributions

$$\begin{cases} \epsilon_i(k) \sim \mathcal{N}(0, \Pi_i(k)) \\ r_i(k) \sim \mathcal{N}(0, R_i(k)) \end{cases} \quad (4.24)$$

where $\Pi_i(k)$ and $R_i(k)$ are to be defined. Note that as a direct consequence of the definition of the estimator in (4.18) and Theorem 4.1 the expected values are $\mathbf{E}[\epsilon_i(k)] = 0$ and $\mathbf{E}[r_i(k)] = 0$, as $\hat{x}_i(k)$ is an unbiased estimate of $x_i(k)$. Indeed, this is also apparent by direct analysis of the dynamics of the estimation error (4.23), considering initialization $\hat{x}_i(0) = \bar{x}_i^0$.

The state estimation error variance matrix $\Pi_i(k) \doteq \text{Var}[\epsilon_i(k)] = \mathbf{E}[\epsilon_i(k)\epsilon_i(k)^\top]$ can be derived from (4.18):

$$\Pi_i(k) = \bar{A}_i(k)A_{ii}\Pi_i(k-1)A_{ii}^\top\bar{A}_i(k)^\top + \bar{A}_i(k)W_i\bar{A}_i(k)^\top + \bar{L}_i(k)V_i\bar{L}_i(k)^\top, \quad (4.25)$$

initialized with $\Pi_i(0) = \Pi_i^0$. Note that the covariance terms $\text{Cov}[\epsilon_i(k-1), w_i(k-1)] = 0$ and $\text{Cov}[\epsilon_i(k-1), v_i(k)] = 0$, given dynamics of $\epsilon_i(k)$ and Assumption 4.3. Moreover, the residual

covariance matrix $R_i(k) = \text{Var}[r_i(k)]$, is derived as:

$$R_i(k) = C_i \Pi_i(k) C_i^\top + V_i - C_i \bar{L}_i(k) V_i - V_i \bar{L}_i(k)^\top C_i^\top, \quad (4.26)$$

where $-C_i \bar{L}_i(k) V_i - V_i \bar{L}_i(k)^\top C_i^\top = \text{Cov}[\epsilon_i(k), v_i(k)]$.

Having presented the statistical characteristics of the estimation error and residual signals, it is important to reiterate that neither can be used directly to detect attacks. Indeed, they are both independent of $x_j(k)$ and $\gamma_{ji}^c(k)$, $j \in \mathcal{N}_i$, and therefore they do not depend on a possible attack on the received communication $\alpha_{ji}(k)$ for any $k \geq k_a^i$. As such, no statistical test based on r_i alone can be used to detect attacks on γ_{ji}^c .

4.2.2 Estimate of physical coupling

Let us now turn our attention to the estimate $\hat{\xi}_i$ of ξ_i defined in (4.14b), and analyse the estimation error associated with it, i.e.

$$\rho_i(k-1|k) \doteq \xi_i(k-1) - \hat{\xi}_i(k-1|k). \quad (4.27)$$

Firstly, given that $\hat{x}_i(k) = x_i(k) - \epsilon_i(k)$ by definition of ϵ_i itself, $\hat{\xi}_i(k-1|k)$ in (4.18) can be rewritten as:

$$\hat{\xi}_i(k-1|k) = \xi_i(k-1) + M_i(k) [C_i (A_{ii} \epsilon_i(k-1) + w_i(k-1)) + v_i(k)] \quad (4.28)$$

implying that the coupling estimation error can be rewritten in terms of local variables only:

$$\rho_i(k-1|k) = -M_i(k) C_i (A_{ii} \epsilon_i(k-1) + w_i(k-1)) - M_i(k) v_i(k). \quad (4.29)$$

Hence, we confirm that $\mathbf{E}[\rho_i(k-1|k)] = 0, \forall k \geq 0$, as stated in Theorem 4.1. Furthermore, it is a realisation of the random distribution

$$\rho_i(k-1|k) \sim \mathcal{N}(0, P_i(k-1|k)), \quad (4.30)$$

with the coupling estimation error variance $P_i(k-1|k) \doteq \text{Var}[\rho_i(k-1|k)]$ evaluated directly from (4.29) as:

$$P_i(k-1|k) = M_i(k) \left(C_i \left(A_{ii} \Pi_i(k-1) A_{ii}^\top + W_i \right) C_i^\top + V_i \right) M_i(k)^\top. \quad (4.31)$$

It must be noted that, given its definition in (4.27), $\rho_i(k-1|k)$ is not directly available to diagnoser $\mathcal{D}_i$. Furthermore, because of the decentralized nature of the estimate, it does not depend on the received information γ_{ji}^c . Thus, similarly to residual r_i , it cannot be used for detection. For this reason, in the following we introduce a second estimate of the aggregate interconnection input ξ_i , providing the necessary redundancy to detect attacks.

Let us define $\check{\xi}_i(k)$ as an estimate of ξ_i computed from received measurements γ_{ji}^c :

$$\check{\xi}_i(k) \doteq G_i^\dagger \sum_{j \in \mathcal{N}_i} A_{ij} \gamma_{ji}^c(k). \quad (4.32)$$

Clearly, in nominal conditions (4.32) can be rewritten as:

$$\check{\xi}_i(k) = G_i^\dagger \sum_{j \in \mathcal{N}_i} A_{ij} \hat{x}_j. \quad (4.33)$$

Thus, we exploit the redundancy between $\hat{\xi}_i$ as defined in (4.18) and $\check{\xi}_i$, to compute the residual-like signal:

$$\hat{\rho}_i(k-1|k) \doteq \hat{\xi}_i(k-1|k) - \check{\xi}_i(k-1), \quad (4.34)$$

which can be seen as a distributed estimate of the coupling estimation error $\rho_i(k-1|k)$. The residual $\hat{\rho}_i$ can then be used to detect attacks, as it is both available to $\mathcal{D}_i$ and dependent on $\alpha_{ji}, j \in \hat{\mathcal{N}}_i$ injected in γ_{ji}^c .

Again, note the one-step time delay introduced by the estimate of the unknown input in (4.18), in turn delaying by one time step the detection of attacks. In Figure 4.3, we summarise the architecture of $\mathcal{O}_i^d$ as described thus far.

Following from definition of $\rho_i(k-1|k)$ in (4.27), in nominal conditions (4.34) can be

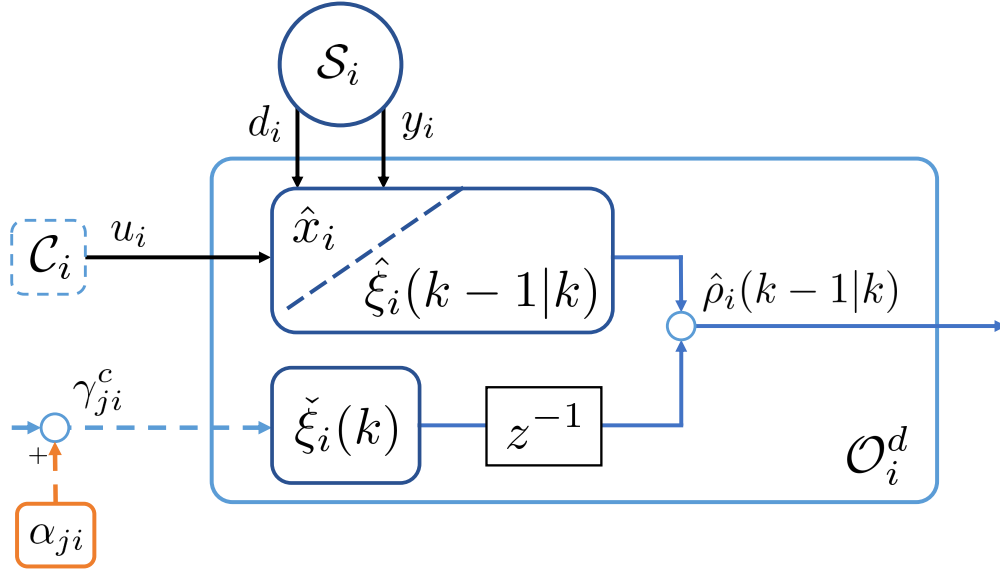


Figure 4.3: Representation of the module $\mathcal{O}_i^d$ within $\mathcal{D}_i$, exploiting trusted signals u_i, y_i and d_i , and received signals γ_{ji}^c . The difference in colour between local signals (in black) and communicated signals (in dashed blue) highlights which are trusted (the former) and which are subject to attacks (the latter). The module is composed of two estimators: one computes $\hat{\xi}_i$ and the other $\tilde{\xi}_i$, both estimating the aggregate physical coupling vector ξ_i (4.12). The time indices of the estimates are included, as well as the one-step time delay block z^{-1} to highlight the fact that the computation of $\tilde{\xi}_i$ requires a delay of one time step. Details are given in Section 4.2.

rewritten as:

$$\begin{aligned} \hat{\rho}_i(k-1|k) &= G_i^\dagger \sum_{j \in \mathcal{N}_i} A_{ij} x_j(k-1) - \rho_i(k-1|k) - \tilde{\xi}_i(k-1) \\ &= G_i^\dagger \sum_{j \in \mathcal{N}_i} A_{ij} [x_j(k-1) - \hat{x}_j(k-1)] - \rho_i(k-1|k) \end{aligned} \quad (4.35)$$

and therefore

$$\hat{\rho}_i(k-1|k) = G_i^\dagger \sum_{j \in \mathcal{N}_i} A_{ij} \epsilon_j(k-1) - \rho_i(k-1|k).$$

The following proposition summarises the statistical properties of $\hat{\rho}_i(k-1|k)$ under nominal conditions.

Proposition 4.1. Consider a subsystem $\mathcal{S}_i$ equipped with diagnoser $\mathcal{D}_i$, including an unbiased minimum-variance filter estimating the local state and physical coupling defined in (4.18). Let Theorem 4.1 hold. The residual $\hat{\rho}_i(k-1|k)$ follows a Gaussian distribution with mean $\mu_i(k-1|k)$

and variance $\Sigma_i(k-1|k)$, where:

$$\mu_i(k-1|k) = 0 \quad (4.36a)$$

$$\Sigma_i(k-1|k) = G_i^\dagger \left(\sum_{j \in \mathcal{N}_i} A_{ij} \Pi_j(k-1) A_{ij}^\top \right) G_i^{\dagger\top} + P_i(k-1|k) \quad (4.36b)$$

during nominal operations, i.e. for all $k \leq \min_{j \in \mathcal{N}_i} k_a^{ji}$, □

Proof. From the definition of $\hat{\rho}_i(k-1|k)$ in nominal conditions (4.35)

$$\mu_i(k-1|k) = \mathbf{E}[\hat{\rho}_i(k-1|k)] = 0,$$

given that $\mathbf{E}[\epsilon_j(k-1)] = 0, \forall j \in \mathcal{N}$, and $\mathbf{E}[\rho_i(k-1|k)] = 0$, for all $k \geq 1$ by construction.

Definition of $\Sigma_i(k-1|k) \doteq \text{Var}[\hat{\rho}_i(k-1|k)]$ in (4.36b) can be derived directly from the definition of variance $\Sigma_i(k-1|k) = \mathbf{E}[\hat{\rho}_i(k-1|k)\hat{\rho}_i(k-1|k)^\top]$ and (4.35), noting that the covariance terms

$$\text{Cov}[\epsilon_j(k-1), \rho_i(k-1|k)] = 0,$$

since the estimation error $\rho_i(k-1|k)$ in $\mathcal{D}_i$ is independent of neighbouring subsystems $\mathcal{S}_j, j \in \mathcal{N}_i$ by construction. ■

Remark 4.6. Note that use of $\Sigma_i(k-1|k)$ is compatible with a distributed and a plug-and-play design of $\mathcal{D}_i$. Indeed, given that all variances Π_i are independent of their neighbors $\mathcal{S}_j, j \in \mathcal{N}_i$, Π_j can be locally computed by $\mathcal{D}_i$ so long as the tuple (A_{jj}, G_j, C_j) and the initial variance Π_j^0 are known for all neighbouring subsystems $\mathcal{S}_j, j \in \mathcal{N}_i$. ◁

Given that $\hat{\rho}_i$ is affected by $\alpha_{ji}, j \in \hat{\mathcal{N}}_i(k) \doteq \{j \in \mathcal{N}_i | k \geq k_a^{ji}, k \in \mathbb{N}_0\}$, it is important to analyse its behaviour also while under attack, to then define an appropriate test to solve Problem 4.1. From the definition of γ_{ji}^c in (4.9), and disregarding the effect of $\beta_{ji}(\cdot)$, the residual $\hat{\rho}_i$ under attack is:

$$\hat{\rho}_i(k-1|k) = G_i^\dagger \left(\sum_{j \in \mathcal{N}_i} A_{ij} \epsilon_j(k-1) - \sum_{j \in \hat{\mathcal{N}}_i(k-1)} A_{ij} \alpha_{ji}(k-1) \right) - \rho_i(k-1|k). \quad (4.37)$$

Thus, during an attack, the mean of $\hat{\rho}_i(k-1|k)$ changes to

$$\mu_i^a(k-1|k) = \sum_{j \in \tilde{\mathcal{N}}_i(k-1)} A_{ij} \alpha_{ji}(k-1), \quad (4.38)$$

while the attack does not have any effect on the variance $\Sigma_i(k-1|k)$, supposing that the attack function $\alpha_{ji}(k) : \mathbb{N}_0 \rightarrow \mathbb{R}^{n_j}$ is deterministic.

4.3 Detection strategy

We are now ready to present an appropriate test to detect cyber-attacks on the communication signals $\gamma_{ji,j \in \mathcal{N}_i}^c$. Let us consider a sliding window of finite length ω_i over which we take the samples of $\hat{\rho}_i(k-1|k)$:

$$\hat{\mathcal{P}}_i(k) \doteq \{\hat{\rho}_i(s-1|s), s = \{k - \omega_i + 1, \dots, k\}\}. \quad (4.39)$$

As we have seen that the residual $\hat{\rho}_i(k-1|k)$ changes mean while maintaining the covariance unaltered when under attack, i.e. for $k > k_a \doteq \min_{j \in \mathcal{N}_i} k_a^{ji}$, we can reformulate the detection problem expressed in Problem 4.1 as the following hypothesis test.

Problem 4.2. Given the collection of samples $\hat{\mathcal{P}}_i$ defined in (4.39), design an appropriate detection test for $\mathcal{D}_i$ such that for each time k one of the following hypotheses

$$\begin{cases} \mathcal{H}_i^0 : & \hat{\rho}_i(k-1|k) \sim \mathcal{N}(\mu_i^h(k-1|k), \Sigma_i(k-1|k)) \\ \mathcal{H}_i^1 : & \hat{\rho}_i(k-1|k) \sim \mathcal{N}(\mu_i^a(k-1|k), \Sigma_i(k-1|k)) \end{cases} \quad (4.40)$$

is accepted, where $\mu_i^h(k-1|k) = 0$ is the nominal mean of $\hat{\rho}_i$, as defined in (4.36a), while μ_i^a is the mean when under attack, as defined in (4.38). $\triangleleft$

Note that Problem 4.2 is equivalent to detecting an unknown deterministic signal in white

Gaussian noise. Indeed, the hypotheses in (4.40) can be reformulated as

$$\begin{cases} \mathcal{H}_i^0 : \hat{\rho}_i(k-1|k) = h_i(k-1) \\ \mathcal{H}_i^1 : \hat{\rho}_i(k-1|k) = \zeta_i^a(k-1) + h_i(k-1) \end{cases} \quad (4.41)$$

where $h_i(k) \sim \mathcal{N}(0, \Sigma_i(k|k+1))$ and the unknown signal to be detected is

$$\zeta_i^a(k) \doteq \sum_{j \in \tilde{\mathcal{N}}_i(k)} A_{ij} \alpha_{ji}(k), \quad (4.42)$$

i.e. the aggregate effect of the attack on the coupling estimate $\check{\xi}_i$.

Having noted this, it is possible to formulate a Generalised Likelihood Ratio Test (GLRT) to distinguish the two hypotheses – see for example [Kay93, BKL⁺06]. The GLRT, following arguments from the Neyman-Pearson Lemma, accepts hypothesis $\mathcal{H}_i^1$ once the ratio of the likelihoods:

$$\frac{p(\hat{\rho}_i(k-1|k) | \hat{\zeta}_i^a(k-\omega_i), \dots, \hat{\zeta}_i^a(k-1), \mathcal{H}_i^1)}{p(\hat{\rho}_i(k-1|k) | \mathcal{H}_i^0)} \quad (4.43)$$

is larger than an appropriately defined threshold, where $\hat{\zeta}_i^a(k)$ is a maximum likelihood estimate of $\zeta_i^a(k)$. Given the definition of the hypotheses in (4.41) and of h_i , $\hat{\rho}_i(k|k+1)$ itself can be such an estimate. Therefore, we define

$$\hat{\zeta}_i^a(k) \doteq \hat{\rho}_i(k|k+1). \quad (4.44)$$

Given that $h_i(k)$ is i.i.d. Gaussian by construction, by taking the natural logarithm of (4.43) over sampling period $\{k-\omega_i+1, \dots, k\}$ it is possible to rewrite (4.43), to obtain the following detection test:

$$\underbrace{\sum_{\kappa=k-\omega_i}^{k-1} \hat{\rho}_i(\kappa|\kappa+1)^\top \Sigma_i(\kappa|\kappa+1)^{-1} \hat{\rho}_i(\kappa|\kappa+1)}_{T(\hat{\rho}_i, k)} > \theta_i(k), \quad (4.45)$$

where the log-likelihood ratio is compared with some appropriately defined detection threshold $\theta_i(k)$. The following holds, making (4.45) well defined.

Lemma 4.1. Consider a subsystem $\mathcal{S}_i$ with dynamics as in (4.5), and a distributed diagnoser

module $\mathcal{D}_i$ equipped with unbiased minimum-variance filter (4.18) estimating the state x_i and the physical coupling ξ_i of $\mathcal{S}_i$. Suppose $M_i(k)$ is defined according to Theorem 4.1, and that $V_i \succ 0$. Then, matrix $\Sigma_i(k|k+1)$ is invertible for all $k \geq 0$. $\square$

Proof. As $\Sigma_i(k|k+1)$ is a covariance matrix, $\Sigma_i(k|k+1) \succcurlyeq 0$ holds by definition, for all $k \geq 0$. Hence, to prove the statement it is sufficient to demonstrate that $\Sigma_i(k|k+1) \succ 0$. Given definition of Σ_i in (4.36b), the definition of $P_i(k|k+1)$ in (4.31), and that $V_i \succ 0$ by Assumption 4.3, the following sufficient condition holds:

$$\text{rank } M_i(k) = g_i \implies P_i(k|k+1) \succ 0 \implies \Sigma_i(k|k+1) \succ 0.$$

Thus, we must show that $M_i(k)$ defined in Theorem 4.1 is full row rank for all $k \in \mathbb{N}_0$. Indeed, from (4.22), $\text{rank } M_i(k) \geq g_i$. Hence, given that $M_i(k) \in \mathbb{R}^{g_i \times p_i}$, its rank satisfies $\text{rank } M_i(k) = \min(g_i, p_i)$. Finally, given that $p_i \geq g_i$ must hold for Assumption 4.7 to be valid, then $\text{rank } M_i(k) = g_i$ is a direct consequence of its definition in Theorem 4.1. Hence $\Sigma_i(k|k+1) \succ 0$ for all $k \geq 0$, and is therefore invertible. $\blacksquare$

From (4.45), given knowledge of the statistical properties of $\hat{\rho}_i(k-1|k)$ described in Proposition 4.1, the probabilities of detection and false alarm can be defined respectively as:

$$\begin{aligned} P_i^f(k) &\doteq \Pr \{T(\hat{\rho}_i, k) > \theta_i(k) | \mathcal{H}_i^0\} \\ P_i^d(k) &\doteq \Pr \{T(\hat{\rho}_i, k) > \theta_i(k) | \mathcal{H}_i^1\} \end{aligned} \quad (4.46)$$

Since $\Sigma_i(k-1|k)$ is symmetric positive definite, there exist a unitary matrix $U_i(k-1|k)$ and a diagonal matrix $\Lambda_i(k-1|k)$ such that

$$\Sigma_i(k-1|k) = U_i(k-1|k) \Lambda_i(k-1|k) U_i(k-1|k)^\top.$$

Thus, by introducing the transformation

$$z_i(k-1|k) \doteq U_i(k-1|k) \hat{\rho}_i(k-1|k),$$

where $z_i \in \mathbb{R}^{g_i}$ is such that all g_i elements are mutually uncorrelated. Furthermore, they are random variables with distribution $z_{i,q} \sim \mathcal{N}(0, \lambda_{i,q})$, where $z_{i,q}$ is the q -th element of z_i and $\lambda_{i,q}$ the q -th diagonal component of Λ_i , for all $q \in \{1, \dots, g_i\}$.

An auxiliary detection statistic $T'(z_{i,q}, k)$ can be defined, thus introducing the detection test:

$$T'(z_{i,q}, k) = \sum_{\kappa=k-\omega_i}^{k-1} \frac{z_{i,q}^2(\kappa|\kappa+1)}{\lambda_{i,q}(\kappa|\kappa+1)} > \vartheta_{i,q}(k) \quad (4.47)$$

where $\vartheta_{i,q}(k) > 0$ is an appropriately defined positive threshold.

Recall that $\hat{\rho}_i(k-1|k) \sim \mathcal{N}(\mu_i(k-1|k), \Sigma_i(k-1|k))$, with

$$\mu_i(k-1|k) \doteq \begin{cases} 0, & \text{for } k \leq k_a; \\ \zeta_i^a(k-1), & \text{otherwise;} \end{cases} \quad (4.48)$$

which follows from definitions of μ_i^h and μ_i^a in (4.36a) and (4.38), respectively. Therefore, given linear relationship between $z_i(k-1|k)$ and $\hat{\rho}_i(k-1|k)$, $z_i(k-1|k)$ has the following distribution:

$$z_i(k-1|k) \sim \mathcal{N}(v_i(k-1|k), \Lambda_i(k-1|k)), \quad (4.49)$$

where the mean $v_i(k-1|k)$ is defined as

$$v_i(k-1|k) \doteq \begin{cases} U_i \mu_i^h(k-1|k) = 0, & \text{for } k \leq k_a \\ U_i \mu_i^a(k-1|k) = U_i \zeta_i^a(k-1), & \text{for } k > k_a \end{cases}. \quad (4.50)$$

Hence, for sufficiently large ω_i , the detection statistic $T'(z_{i,q}, k)$ is a random variable with distribution:

$$T'(z_{i,q}, k) \sim \begin{cases} \chi_{\omega_i}^2(0), & \text{if } \mathcal{H}_i^0 \\ \chi_{\omega_i}^2(\nu_{i,q}(k)), & \text{if } \mathcal{H}_i^1 \end{cases} \quad (4.51)$$

where with $\chi_{\omega_i}^2(c)$ we define a chi-squared distribution with degrees of freedom ω_i and non-

centrality parameter c . Here, the non-centrality parameter $\nu_{i,q}(k)$ is defined as:

$$\begin{aligned}\nu_{i,q}(k) &\doteq \sum_{\kappa=k-\omega_i}^{k-1} \frac{1}{\lambda_{i,q}(\kappa|\kappa+1)} v_{i,q}^2(\kappa|\kappa+1) \\ &= \sum_{\kappa=k-\omega_i}^{k-1} \frac{1}{\lambda_{i,q}(\kappa|\kappa+1)} (U_{i,[q]}(\kappa|\kappa+1) \zeta_i^a(\kappa))^2\end{aligned}\quad (4.52)$$

with $U_{i,[q]}$ the q -th row of U_i .

Finally, defining the tail probability of the normalised χ^2 distribution for $T'(z_{i,q}, k)$ as

$$\Phi(\mathbf{x}) \doteq \Pr\{T'(z_{i,q}, k) > \mathbf{x}\} \quad (4.53)$$

for some threshold $\mathbf{x}$, it is possible to compute the false alarm and detection probabilities for test (4.51) for a given threshold $\vartheta_{i,q}$. Indeed, they can be computed by evaluating the following [Kay93, Appendix 7.A] for each component $q = \{1, \dots, g_i\}$:

$$P_{i,q}^f(k) = \Phi\left(\frac{1}{\sqrt{2\omega_i}} (\vartheta_{i,q}(k) - \omega_i)\right) \quad (4.54)$$

$$P_{i,q}^d(k) = \Phi\left(\frac{1}{\sqrt{4\nu_{i,q}(k) + 2\omega_i}} (\vartheta_{i,q}(k) - \omega_i - \nu_{i,q}(k))\right). \quad (4.55)$$

Note furthermore that (4.55) can be usefully rewritten as a function of the false alarm probability, as

$$P_{i,q}^d = \Phi\left(\frac{1}{\sqrt{4\nu_{i,q}(k) + 2\omega_i}} \left(\sqrt{2\omega_i} \Phi^{-1}(P_{i,q}^f(k)) - \nu_{i,q}(k)\right)\right) \quad (4.56)$$

Following this the detection threshold $\vartheta_{i,q}$ can be defined such that an acceptable false alarm rate $P_i^f(k)$ is achieved. Firstly, having defined the probability of false alarm component-wise, the overall probability of false alarm of detector $\mathcal{D}_i$ can be expressed as:

$$P_i^f(k) \doteq 1 - \prod_{q=1}^{g_i} (1 - P_{i,q}^f(k)). \quad (4.57)$$

In fact, given that the components of $z_i(k)$ are independent by construction, the probability of at least one false alarm is complementary to the probability of no false alarms, leading to (4.57). By selecting $P_{i,q}^f(k)$ common to all $q = \{1, \dots, g_i\}$, it is then possible to find $\vartheta_{i,q}$ by

inverting (4.57) and (4.54). For ω_i sufficiently large, the same can be done to determine the overall probability of detection $P_i^d(k)$ for a signal of given average energy.

Remark 4.7. The GLRT presented here can be interpreted as an Energy Detector (see [Kay93, Ch. 7]), i.e. a Neyman-Pearson detector that computes the energy of the signal and compares it to a threshold. Indeed, $T'(z_{i,q}, k)$ can be seen as the energy of the cumulative attack on the communication signals received by $\mathcal{S}_i$. In fact, the probability of detection increases with the energy of the attack, as $\nu_{i,q}$ is proportional to attack's energy scaled by the variance of $z_{i,q}$, while for $\nu_{i,q} \rightarrow 0$ the probability of detection approaches that of false alarm.

In addition, note that for all time-varying attacks, $z_{i,q}(\kappa|\kappa+1)$ appearing in $T'(z_{i,q}, k)$ may not be independent Gaussian processes over the interval $\kappa \in \{k - \omega_i, \dots, k - 1\}$. As such, the probability of detection in (4.55) holds only for sufficiently large ω_i [Kay93, Appedix 7A], as under this condition the detection test $T'(z_{i,q}, k)$ can be adequately described as the non-central χ^2 distribution with ω_i degrees of freedom given in (4.51). On the other hand, (4.54) holds independently of ω_i , as in nominal conditions $z_{i,q}(k|k+1)$ are independent zero mean Gaussian processes. This is important, as $P_{i,d}^f$ is used to compute the detection threshold $\vartheta_{i,q}$. $\triangleleft$

We have now presented the detection test available to $\mathcal{D}_i$, presenting its statistical properties and how it is distributed and scalable, provided model information (A_{jj}, G_j, C_j) and Π_j^0 available to $\mathcal{D}_i$ for all $j \in \mathcal{N}_i, i \in \mathcal{N}$, as it depends only on local information on $\mathcal{S}_i$ and on information of its neighbours.

It is worth pointing out the following:

Proposition 4.2. Consider a subsystem equipped with a diagnoser $\mathcal{D}_i$ with decentralized state and coupling estimator in (4.18), and detection test (4.51). If, for all $k \geq k_a$, attack signals $\alpha_{ji}(k) \neq 0, j \in \widehat{\mathcal{N}}_i(k)$, satisfy

$$\text{col}_{j \in \widehat{\mathcal{N}}_i(k)} (\alpha_{ji}(k)) \in \ker \left[\text{row}_{j \in \widehat{\mathcal{N}}_i} (A_{ij}) \right], \quad (4.58)$$

then the attack is stealthy to the diagnoser $\mathcal{D}_i$. $\square$

Proof. The proof follows directly from the definition of γ_{ji}^c and (4.42). $\blacksquare$

Proposition 4.2 demonstrates that the diagnosis module $\mathcal{D}_i$ designed up until this point is still subject to the structurally stealthy attacks described in Chapter 3.2. This is a clear consequence of the decentralized design of the unbiased minimum-variance filter in (4.18). Indeed, although (A_{jj}, G_j, C_j) and Π_j^0 are necessary for the computation of $\Sigma_i(k|k+1)$, and are therefore included in the information-set of $\mathcal{O}_i^d$, they are not exploited to directly monitor γ_{ji}^c . In the following Section, we describe a distributed filter for the estimation of the states of neighbouring subsystems $\mathcal{S}_j, j \in \mathcal{N}_i$ in module $\mathcal{O}_{ji}^d$.

4.4 Detection filter for neighbouring subsystems

Having presented the local module $\mathcal{O}_i^d$ composed of the decentralized filter (4.18), residual generator (4.34) and detection test (4.51), in the remainder of this chapter we focus on the design of the additional filters $\mathcal{O}_{ji}^d$. These modules are designed to estimate the state of all subsystems $\mathcal{S}_j, j \in \mathcal{N}_i$, to dynamically monitor the trajectory of γ_{ji}^c . Exploiting available knowledge of the dynamics of the neighbouring subsystem, we design a decentralized estimate $\hat{x}_{ji}$ via an unbiased minimum-variance filter.

Let us first redefine γ_{ji}^c from its definition in (4.9) to

$$\gamma_{ji}^c(k) = y_j(k) + \beta_{ji}(k - k_a^{ji})\alpha_{ji}(k), \quad (4.59)$$

where y_j is the measurement output of $\mathcal{S}_j$, α_{ji} is the attack signal, and $\beta_{ji}(\cdot)$ is the activation function, as defined in (4.10).

Remark 4.8. We redefine the communicated variable γ_{ji}^c , as the main objective of this section is to design a filter to estimate the state of the neighbouring subsystems via an estimator $\hat{x}_{ji}, j \in \mathcal{N}_i$.

However, it is possible that γ_{ji}^c as defined in (4.59) is inappropriate given the control law $\kappa_i(\cdot, \cdot)$ in (4.8), i.e. $\kappa_i(y_i, C_j^\dagger \gamma_{ji, j \in \mathcal{N}_i})$ may not achieve the desired objective. In this case, it is sufficient to change the definition of the control input u_i in $\mathcal{C}_i$ to

$$u_i = \kappa_i(y_i, \hat{x}_{ji, j \in \mathcal{N}_i}). \quad (4.60)$$

Furthermore, the redefinition of γ_{ji}^c calls for changes in our definition of the coupling estimate $\check{\xi}_i$ (4.32), and therefore for changes to the statistical properties of the residual $\hat{\rho}_i$ (4.34). The estimate $\check{\xi}_i$ may be redefined in two different ways: if C_j is such that

$$\ker C_j \subseteq \ker A_{ij} \quad (4.61)$$

holds for all neighbours, then the estimate's definition can be changed to:

$$\check{\xi}_i = G_i^\dagger \sum_{j \in \mathcal{N}_i} A_{ij} C_j^\dagger \gamma_{ji}^c. \quad (4.62)$$

This in turn implies that the statistical properties of $\hat{\rho}_i(k-1|k)$ must change: specifically, covariance matrix Σ_i must be redefined from (4.36b) to

$$\Sigma_i(k-1|k) = G_i^\dagger \left(\sum_{j \in \mathcal{N}_i} A_{ij} C_j^\dagger V_j C_j^{\dagger\top} A_{ij}^\top \right) G_i^{\dagger\top} + P_i(k-1|k) \quad (4.63)$$

with V_j defined in (4.6) and $P_i(k-1|k)$ defined in (4.31).

On the other hand, if (4.61) does not hold, we can rely once again on the local construction of $\hat{x}_{ji}$ to compute $\check{\xi}_i$, similarly to the argument in Remark 3.11. Indeed, redefining the estimate of the aggregate coupling as

$$\check{\xi}_i = G_i^\dagger \sum_{j \in \mathcal{N}_i} A_{ij} \hat{x}_{ji}, \quad (4.64)$$

note that it is unbiased, following from unbiasedness of $\hat{x}_{ji}$ [Theorem 4.1]. Hence the estimation error $\hat{\rho}_i(k-1|k)$ has zero mean, and its covariance matrix is

$$\Sigma_i = G_i^\dagger \left(\sum_{j \in \mathcal{N}_i} A_{ij} \Pi_{ji}(k-1) A_{ij}^\top \right) G_i^{\dagger\top} + P_i(k-1|k), \quad (4.65)$$

where $\Pi_{ji}(k-1)$ is the covariance matrix of the estimation error $\epsilon_{ji} \doteq x_j - \hat{x}_{ji}$, defined in the following. No other changes to the diagnostic module $\mathcal{O}_i^d$ are necessary⁵. $\triangleleft$

Suppose Assumptions 4.4 and 4.7 hold. Note that, for the local filter (4.18), knowledge of

⁵Recall that Σ_i is used for the computation of the statistical test $T(\hat{\rho}_i, k)$ in (4.45), and therefore that these changes do affect the rest of the diagnostic module, even though definitions do not have to change.

u_i and d_i influencing the dynamics of $\mathcal{S}_i$ (4.5) is assumed to be available to $\mathcal{D}_i$ when computing $\hat{x}_i$. For the estimate of x_j by $\mathcal{D}_i$, however, this can no longer be the case, as including them in the communication vector γ_{ji}^c may not be possible, either for restrictions on the communication resources or for privacy. Additionally, this information itself may be exposed to attacks, thus requiring further monitoring.

Hence, we rewrite the dynamics of $\mathcal{S}_j$ in (4.5) as:

$$\begin{cases} x_j(k+1) = A_{jj}x_j(k) + \bar{G}_j\bar{\xi}_j(k) + w_j(k) \\ y_j(k) = C_jx_j(k) + v_j(k) \end{cases} \quad (4.66)$$

where we have introduced a new unknown-input vector $\bar{\xi}_j \in \mathbb{R}^{\bar{g}_j}$ satisfying

$$\bar{G}_j\bar{\xi}_j \doteq B_ju_j + E_jd_j + G_j\xi_j, \quad (4.67)$$

with ξ_j defined in (4.12), and matrix $\bar{G}_j$ such that it is a basis for $\text{Im}[B_j, E_j, G_j]$. Thus, $\bar{G}_j \in \mathbb{R}^{n_j \times \bar{g}_j}$ is full column rank.

Assumption 4.8. *Output matrices C_i are such that*

$$\text{rank}(C_i\bar{G}_i) = \text{rank}(\bar{G}_i) = \bar{g}_i, \quad (4.68)$$

for all subsystems $\mathcal{S}_i$, $i \in \mathcal{N}$. ◁

The neighbouring state estimate $\hat{x}_{ji}(k)$ computed by $\mathcal{O}_{ji}^d$ is built using a similar filter to that presented in Section 4.2, with the necessary changes made to account for the different unknown input vector. Specifically, it is the resultant of the following dynamical system:

$$\hat{x}_{ji}(k|k-1) = A_{jj}\hat{x}_{ji}(k-1|k-1) \quad (4.69a)$$

$$\hat{x}_{ji}^*(k|k) = \hat{x}_{ji}(k|k-1) + \bar{G}_jM_{ji}(k) [\gamma_{ji}^c(k) - C_{ji}\hat{x}_{ji}(k|k-1)] \quad (4.69b)$$

$$\hat{x}_{ji}(k|k) = \hat{x}_{ji}^*(k|k) + K_{ji}(k) [\gamma_{ji}^c(k) - C_{ji}\hat{x}_{ji}^*(k|k)] \quad (4.69c)$$

which, omitting the conditional time dependencies, is equivalent to:

$$\hat{x}_{ji}(k) = \bar{A}_{ji}(k)A_{jj}\hat{x}_{ji}(k-1) + \bar{L}_{ji}(k)\gamma_{ji}^c(k) \quad (4.70)$$

with $\bar{M}_{ji}(k)$, $\bar{K}_{ji}(k)$, $\bar{A}_{ji}(k)$, and $\bar{L}_{ji}(k)$ defined as in (4.15), (4.16), and (4.20), respectively.

Note that an appropriate estimate $\hat{\xi}_{ji}$ of $\bar{\xi}_j$ could also be computed in (4.69); however, it is not considered in (4.69), as it is not useful for the detection of attacks on γ_{ji}^c by $\mathcal{O}_{ji}^d$. Once again, the estimate $\hat{x}_{ji}(k)$ is only computable once $\gamma_{ji}^c(k)$ is received, to account for the unknown effect of $\bar{\xi}_j(k-1)$ on $x_j(k)$.

Following the analysis throughout Section 4.2, and specifically Theorem 4.1, we know that the estimate $\hat{x}_{ji}$ is unbiased in nominal conditions, and that its estimation error $\epsilon_{ji} \doteq x_j - \hat{x}_{ji}$ and the residual $r_{ji} \doteq \gamma_{ji}^c - C_j\hat{x}_{ji}$ are subject to the dynamics

$$\begin{cases} \epsilon_{ji}(k) = \bar{A}_{ji}(k)[A_{jj}\epsilon_{ji}(k-1) + w_j(k-1)] - \bar{L}_{ji}(k)v_j(k) \\ r_{ji}(k) = C_j\epsilon_{ji}(k) + v_j(k). \end{cases} \quad (4.71)$$

Furthermore, given Assumption 4.3, ϵ_{ji} and r_{ji} are two random variables with Gaussian distributions:

$$\begin{cases} \epsilon_{ji}(k) \sim \mathcal{N}(0, \Pi_{ji}(k)) \\ r_{ji}(k) \sim \mathcal{N}(0, R_{ji}(k)) \end{cases} \quad (4.72)$$

where the covariance matrices $\Pi_{ji}(k)$ and $R_{ji}(k)$ are defined respectively by:

$$\Pi_{ji}(k+1) = \bar{A}_{ji}(k)A_{jj}\Pi_{ji}(k)A_{jj}^\top\bar{A}_{ji}(k)^\top + \bar{A}_{ji}(k)W_i\bar{A}_{ji}(k)^\top + \bar{L}_{ji}(k)V_j\bar{L}_{ji}(k)^\top, \quad (4.73)$$

$$R_{ji}(k) = C_j\Pi_{ji}(k)C_j^\top + V_j - C_j\bar{L}_{ji}(k)V_j - V_j\bar{L}_{ji}(k)^\top C_j^\top. \quad (4.74)$$

Remark 4.9. Given the definition of $M_{ji}(k)$ and $K_{ji}(k)$ in (4.15)-(4.16), the residual covariance $R_{ji}(k)$ is not invertible, i.e. $R_{ji}(k) \succcurlyeq 0$. This follows directly from the design of estimation filter (4.69) such that the estimate $\hat{x}_{ji}(k)$ is unbiased by unknown input $\bar{\xi}_j(k-1)$. Specifically, it can be verified that $R_{ji}(k)$ has $\bar{g}_i$ eigenvalues at 0, implying that $\bar{g}_i$ components of $r_{ji}(k)$ are a linear combination of the remaining $p_j - \bar{g}_i$ components. This can be understood intuitively, as

the filter computing $\hat{x}_{ji}(k)$ must decouple the estimate from a $\bar{g}_i$ -dimensional subspace of $\mathbb{R}^{n_j}$, which corresponds to the reachability subspace of $\bar{\xi}_i(k-1)$ after a single time step. Therefore, the estimated state is a linear combination of the $p_j - \bar{g}_j$ -dimensional subspace of $\mathbb{R}^{p_j}$ unaffected by $\bar{\xi}_i$. $\triangleleft$

The fact that $R_{ji}(k) \succcurlyeq 0$ is fundamental for the design of a detection test $T(r_{ji}, k)$, designed to detect attacks on γ_{ji}^c , as is shown in the following. Indeed, recalling definition of $T(\hat{\rho}_i, k)$ in (4.45), the inverse of the covariance of $\hat{\rho}_i$ was used as a scaling factor.

Before going on to define the statistical test to detect attacks from $r_{ji}(k)$, however, let us analyse the behaviour of the estimation filter while under attack. Consider that for some time $k_a > 0$ a malicious agent leverages an attack on γ_{ji}^c , for some $j \in \mathcal{N}_i$. Consequently, $\hat{x}_{ji}$ is influenced by the attack for $k \geq k_a$, in turn influencing the estimation error and residual.

To better analyse the behaviour of the estimation error and of the residual under attack, let us divide ϵ_{ji} and r_{ji} in their *healthy* and *attacked* components, i.e.:

$$\begin{aligned}\epsilon_{ji}(k) &= \epsilon_{ji}^h(k) + \epsilon_{ji}^a(k) \\ r_{ji}(k) &= r_{ji}^h(k) + r_{ji}^a(k)\end{aligned}\tag{4.75}$$

where the healthy components behave according to the nominal dynamics (4.71). On the other hand, exploiting linearity, the attacked components have dynamics⁶:

$$\begin{aligned}\epsilon_{ji}^a(k) &= \bar{A}_{ji}(k)A_{jj}\epsilon_{ji}^a(k-1) - \bar{L}_{ji}\alpha_{ji}(k) \\ r_{ji}^a(k) &= C_j\epsilon_{ji}^a(k) + \alpha_{ji}(k).\end{aligned}\tag{4.76}$$

The statistical properties of the estimation error and the residual under attack can therefore be derived, for $k \geq k_a$, as

$$\begin{aligned}\epsilon_{ji}(k) &\sim \mathcal{N}(\epsilon_{ji}^a(k), \Pi_{ji}(k)) \\ r_{ji}(k) &\sim \mathcal{N}(r_{ji}^a(k), R_{ji}(k))\end{aligned}\tag{4.77}$$

given that $\alpha_{ji}(k)$ is considered to be a deterministic signal.

⁶As previously, we here disregard the effect of scaling function $\beta_{ji}(\cdot)$ for sake of clarity of presentation.

Similarly to Section 4.3, resolving Problem 4.1 is equivalent to detecting the presence of a deterministic signal $r_{ji}^a(k)$ in white Gaussian noise. Hence, by exploiting a GLRT, we can define an appropriate statistical test, and then compare it to a threshold to detect the attack. Specifically, recall that $R_{ji} \succcurlyeq 0$ with $\bar{g}_i$ eigenvalues at 0, and the definition of $T'(z_i, k)$ in (4.47). Thus, introducing a window of length ω_{ji} , we define the following component-by-component statistical test, for all $q \in \mathcal{Q}_j \doteq \{1, \dots, p_j - \bar{g}_j\}$:

$$T'(\phi_{ji,q}, k) = \sum_{\kappa=k-\omega_{ji}+1}^k \frac{\phi_{ji,q}(\kappa)^2}{\psi_{ji,q}(\kappa)} > \bar{\vartheta}_{i,q}(k) \quad (4.78)$$

where $\phi_{ji,q}(k)$ is the q -th component of $\phi_{ji}(k)$, itself an auxiliary variable, defined as

$$\phi_{ji}(k) = \Upsilon_{ji}(k)r_{ji}(k), \quad (4.79)$$

with $\Upsilon_{ji}(k)$ a unitary matrix such that

$$R_{ji}(k) = \Upsilon_{ji}(k)\Psi_{ji}(k)\Upsilon_{ji}(k)^\top. \quad (4.80)$$

Existence of $\Upsilon_{ji}(k)$ follows directly from symmetry of R_{ji} . Note that in (4.78) we consider $q \in \mathcal{Q}_j$, supposing without loss of generality that $\psi_{ji,q} \neq 0$ for all $q \in \mathcal{Q}_j$, with $\psi_{ji,q}(k)$ the q -th diagonal component of $\Psi_{ji}(k)$.

Given linear relationship between $r_{ji}(k)$ and $\phi_{ji}(k)$, the latter is a Gaussian random variable with distribution:

$$\phi_{ji}(k) \sim \mathcal{N}(\varsigma_{ji}(k), \Psi_{ji}(k)). \quad (4.81)$$

Similarly to (4.50), the definition of the mean $\varsigma_{ji}(k)$ changes depending on whether the communication signal γ_{ji}^c is attacked or not. Indeed, recalling the definition of the residual under attack $r_{ji}^a(k)$ in (4.76), the expected value of $\phi_{ji}(k)$ can be defined as:

$$\varsigma_{ji}(k) \doteq \begin{cases} 0, & \text{if } k < k_a; \\ \Upsilon_{ji}r_{ji}^a(k), & \text{otherwise.} \end{cases} \quad (4.82)$$

This in turn implies that the test $T'(\phi_{ji,q}, k)$ can be described, for sufficiently large sampling window ω_{ji} , and for each component q , as a random variable with χ^2 distribution with the following characteristics:

$$T'_i(\phi_{ji,q}, k) \sim \begin{cases} \chi_{\omega_{ji}}^2(0), & \text{if } k, k_a; \\ \chi_{\omega_{ji}}^2(\nu_{ji,q}(k)), & \text{otherwise} \end{cases} \quad (4.83)$$

with ω_{ji} degrees of freedom and non-centrality parameter $\nu_{ji}(k)$:

$$\nu_{ji,q}(k) \doteq \sum_{\kappa=k-\omega_{ji}+1}^k \frac{1}{\psi_{ji,q}} (\Upsilon_{ji,[q]}(\kappa) r_{ji}^a(\kappa)), \quad (4.84)$$

with $\Upsilon_{ji,[q]}(k)$ the q -th row of $\Upsilon_{ji}(k)$.

Finally, given distribution (4.83), the probability of detection and false alarm of the diagnoser module $\mathcal{O}_{ji}^d$ can be computed [Kay93, Chapter 7.A]

$$P_{ji,q}^f(k) = \Phi \left(\frac{1}{\sqrt{2\omega_{ji}}} (\vartheta_{ji,q}(k) - \omega_{ji}) \right) \quad (4.85)$$

$$P_{ji,q}^d(k) = \Phi \left(\frac{\sqrt{2\omega_{ji}} \Phi^{-1}(P_{ji,q}^f(k)) - \nu_{ji,q}}{\sqrt{4\nu_{ji,q} + 2\omega_{ji}}} \right) \quad (4.86)$$

where $\Phi(\cdot)$ is the tail probability of the χ^2 distribution with ω_{ji} degrees of freedom, defined in (4.53).

Before concluding this section, it is important to note how, similarly to the module $\mathcal{O}_{ji}^{UIO}$ in Chapter 3, $\mathcal{O}_{ji}^d$ is comparable to a monitoring scheme for stand-alone subsystem $\mathcal{S}_j$. Indeed, it only relies on model knowledge of $\mathcal{S}_j$ and on received output measurements γ_{ji}^c to detect attacks, treating ξ_j, u_j and d_j as unknown disturbances. As such, as shown in the previous chapter for the deterministic case, it is not capable of detecting any of the classes of attacks stealthy to centralized diagnosers, as summarized in Chapter 2.

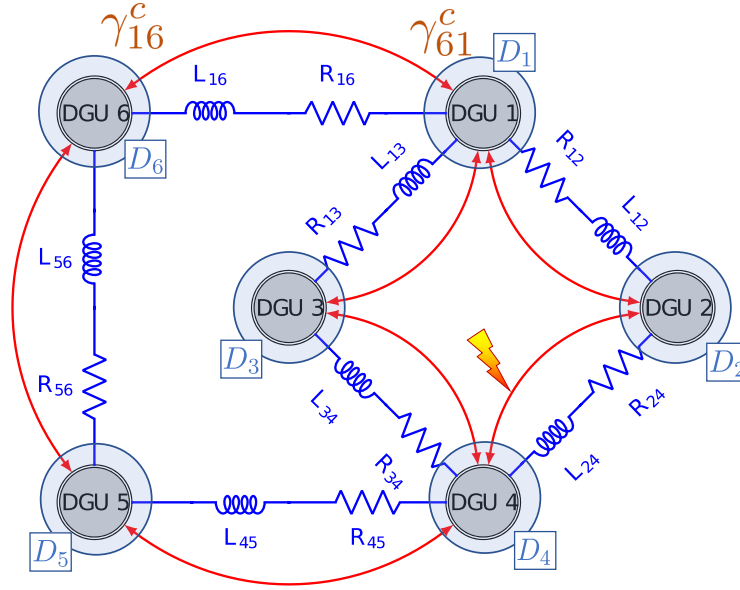


Figure 4.4: DC microgrid represented as a LSS with $N = 6$ DGUs as subsystems. Each node represents a DGU, in blue the physical power lines between subsystems, in red the bi-directional communication links. All DGUs are equipped with a diagnoser $\mathcal{D}_i$.

Table 4.2: Electrical parameters of the DC microgrid

Converter parameters				Power line parameters		
DGU	R_{ti} [Ω]	L_{ti} [mH]	C_{ti} [mF]	Connected DGUs (j, i)	R_{ij} [Ω]	L_{ij} [μ H]
$\mathcal{S}_1$	0.2	1.8	2.2	(1, 2)	0.05	2.1
$\mathcal{S}_2$	0.3	2.0	1.9	(1, 3)	0.07	1.8
$\mathcal{S}_3$	0.1	2.2	1.7	(1, 6)	0.1	2.5
$\mathcal{S}_4$	0.5	3.0	2.5	(2, 4)	0.04	2.3
$\mathcal{S}_5$	0.4	1.3	2.0	(3, 4)	0.06	1.0
$\mathcal{S}_6$	0.6	2.5	3.0	(5, 4)	0.08	1.8
				(5, 6)	0.08	3.0

4.5 Simulation results – DC microgrids

Let us now demonstrate the effectiveness of the proposed strategy through simulation. We consider a similar scenario as that in Chapter 3.9: we take a network of $N = 6$ DGUs, interconnected as in Figure 4.4, with parameters defined in Table 4.2. We refer the reader to Chapter 2 for a detailed overview of the dynamical modelling of DC microgrids, and to Appendix B for a detailed analysis of their discretisation.

Following the discussion in the introduction to this chapter, we suppose that the discretisation of the DC microgrid is performed subsystem-by-subsystem, thus preserving the interconnection structure of the LSS, and that it is done using zero-order hold discretisation method for the local signals, considering neighbouring states $x_j, j \in \mathcal{N}_i$ as exogenous inputs, as discussed in Appendix B. We consider the sampling frequency to be the same as the switching frequencies of the Buck converters in the DGUs, i.e. $T_s = T_i^{\text{sw}} = 0.1\text{ms}$, for all $i \in \mathcal{N}$, and that all sampling be synchronised.

Similarly to Chapter 3.9, we must include augmented dynamics for each DGU, as presented in Appendix C. Indeed, recall that for each $\mathcal{O}_{ji}^d$ the unknown input is

$$\bar{\xi}_j = \bar{G}_j^\dagger (B_j u_j + E_j d_j + G_j \xi_j) \in \mathbb{R}^{\bar{g}_j}.$$

From definition of DGU dynamics in Chapter 2, $\bar{g}_i = n_i, \forall i \in \mathcal{N}$. Thus R_{ji} cannot be used for detection. By including the augmented dynamics defined in Appendix C, however, $n_i = 6$ for all $\mathcal{S}_i$, while $\bar{g}_i = 4$. Thus, $\text{rank } R_{ji} = 2$ for all $j \in \mathcal{N}_i, i \in \mathcal{N}$, allowing detection.

The process and measurement noise covariance matrices are defined for the actual dynamics of the DGU (as opposed to the augmented dynamics) as

$$W_i \doteq 10^{-10} I_3, \quad V_i \doteq \begin{bmatrix} 10^{-4} & & \\ & 10^{-4} & \\ & & 10^{-14} \end{bmatrix}, \quad (4.87)$$

for all $i \in \mathcal{N}$. Note that these values are consistent with those taken into consideration in the previous chapter, given discretisation. Furthermore, it is reasonable to consider $\tilde{V}_{i,3}$ to be considerably smaller than measurement noise on other variables, as $x_{i,3} = \nu_i$ is a controller state internal to $\mathcal{C}_i$, and thus not subject to measurement error. Similarly, it is considered that the augmented state is not subject to significant disturbances.

For the simulation, we consider false alarm values of $P_i^f = P_{ji,q}^f = 5\%$ for all components q , and windows $\omega_i = 1$ and $\omega_{ji} = 5$. By setting $\omega_i = 1$, we are making the test $T'(z_i, k)$ more susceptible to transients, while with $\omega_{ji} = 5$, the detection test is more sensible to steady state

changes.

We propose the following attack scenario: considering a coordinated attack on γ_{24}^c and γ_{34}^c starting from $t = 8$ s, we suppose that α_{24} and α_{34} are both covert attacks. Specifically, we consider that the attacks can simulate the dynamics of DGU 2 and DGU 3 when connected to their neighbours, as well as their augmented dynamics.

We suppose that the attacker simulates Δx_{24}^a and Δx_{34}^a with dynamics as in (3.95), and with exogenous input

$$\Delta d_{ji}^a = \begin{bmatrix} 0 \\ \Delta V_{ref,j} \end{bmatrix}, \quad (4.88)$$

i.e. simulating a change in their local voltage reference, with $i = 4$ and $j \in \{2, 3\}$. This is consistent with a change in the secondary control layer of DGU 2 and DGU 3, which may be caused by a load change in their neighbours.

In the attack scenario we consider, we suppose that for a period of $T_I = 2$ s, $\Delta V_{ref,j} = -0.15$ V for both α_{24} and α_{34} . After this, for $t \geq 10$ s, $\Delta V_{ref,j} = 0$ V, $j \in \{2, 3\}$. Thus Δx_{24}^a and Δx_{34}^a are controlled to the stable equilibrium at the origin. The decision to have the attacks gradually transition back to zero was made to avoid transients. Indeed, if the attacks were to abruptly stop at $T_a + T_I$, i.e. $\alpha_{24}(t) = \alpha_{34}(t) = 0$ for $t \geq 10$ s, the transient caused would bring $\mathcal{O}_{24}^d$ and $\mathcal{O}_{34}^d$ to detect the attacks. We present the results in Figure 4.5 and Figure 4.6.

In Figure 4.5 we show how the local diagnosis module $\mathcal{O}_4^d$ detects the attack at time $t = 8.4570$ s. In Figure 4.6, on the other hand, we show that a covert attack is indeed stealthy to $\mathcal{O}_{24}^d$ and $\mathcal{O}_{34}^d$, estimating the state of the neighbouring dynamics. Finally, in Figure 4.7 we show the impact of the attack on the DC microgrid's states. Over the course of the attack, it is clear that the consensus algorithm is interrupted, while the average voltage of the DC microgrid drops by approximately 1.5 V. It is important to note that, once the attack is interrupted, the current sharing returns to a consensus equilibrium, while the average voltage fails to track $V_{ref} = 48$ V.

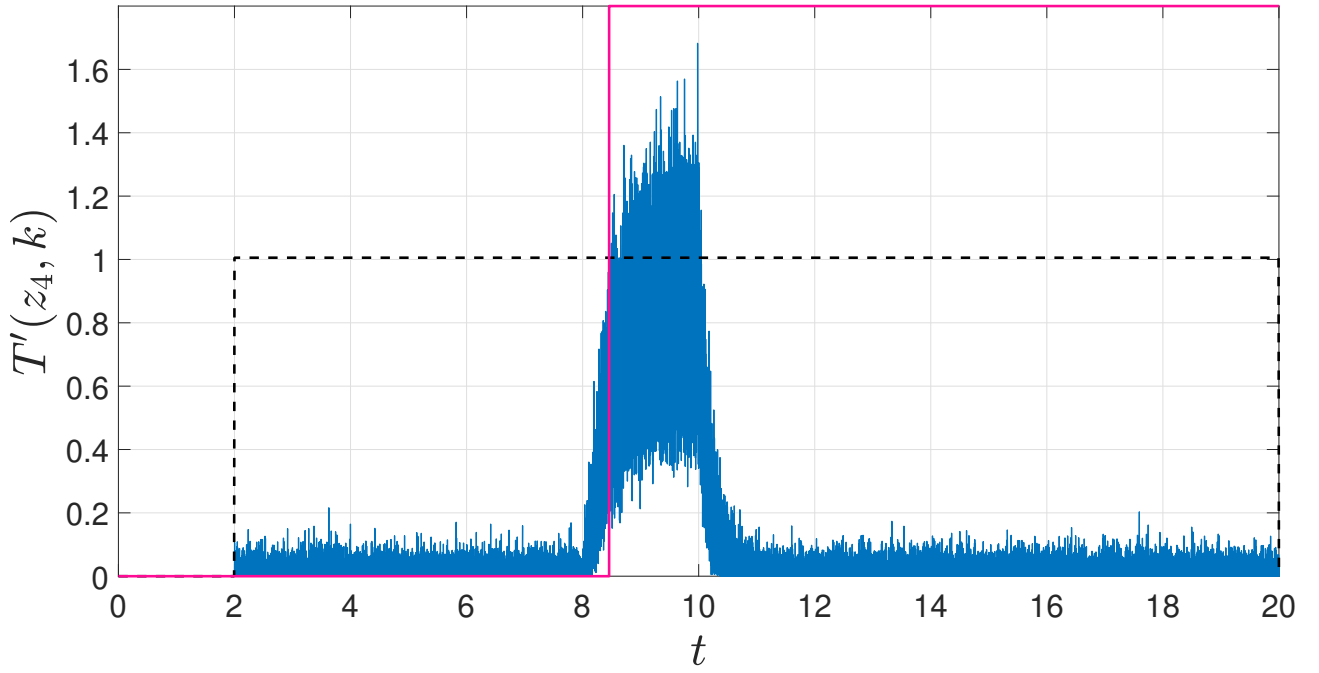
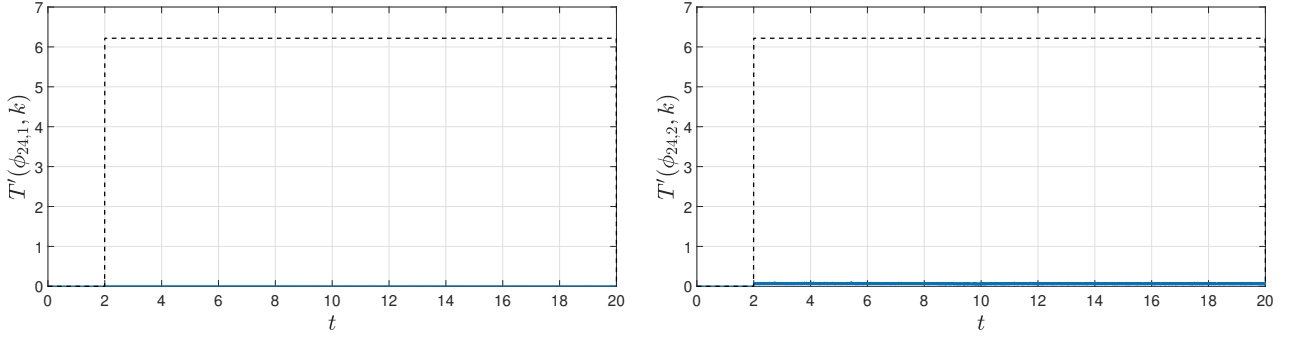
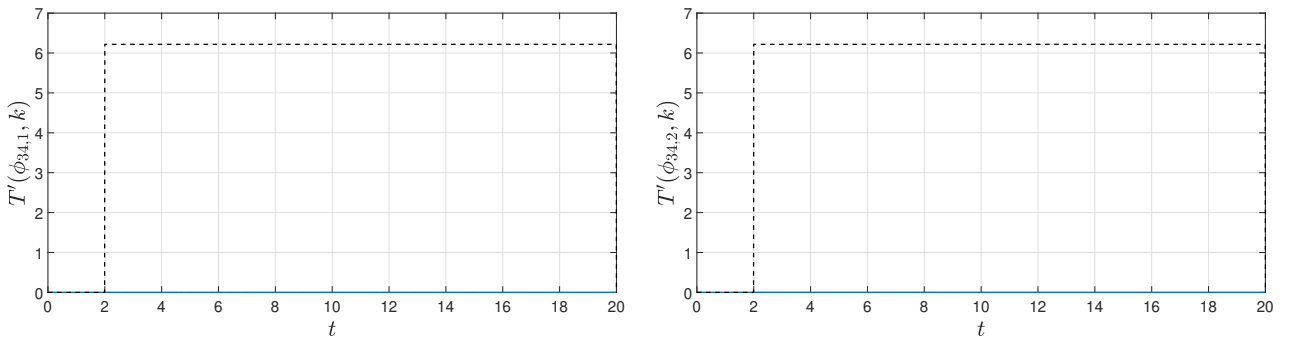


Figure 4.5: Comparison of the test statistic $T'(z_4, k)$, in blue, against the detection threshold $\bar{\vartheta}_4(k)$, in dashed-black. A flag indicating when attack is detected is given in pink.



(a) Detection test and threshold, $\mathcal{O}_{24}^d$



(b) Detection test and threshold, $\mathcal{O}_{34}^d$

Figure 4.6: Comparison of the statistic $T'(\phi_{j4,q}, k)$, in blue, against the detection threshold $\bar{\vartheta}_{j4,q}(k)$, in dashed-black, for $j \in \{2, 3\}$ and $q \in \{1, 2\}$. As expected, the covert attack does not have an effect on the residuals r_{24} or r_{34} , and therefore is not detected by either $\mathcal{O}_{24}^d$ or $\mathcal{O}_{34}^d$.

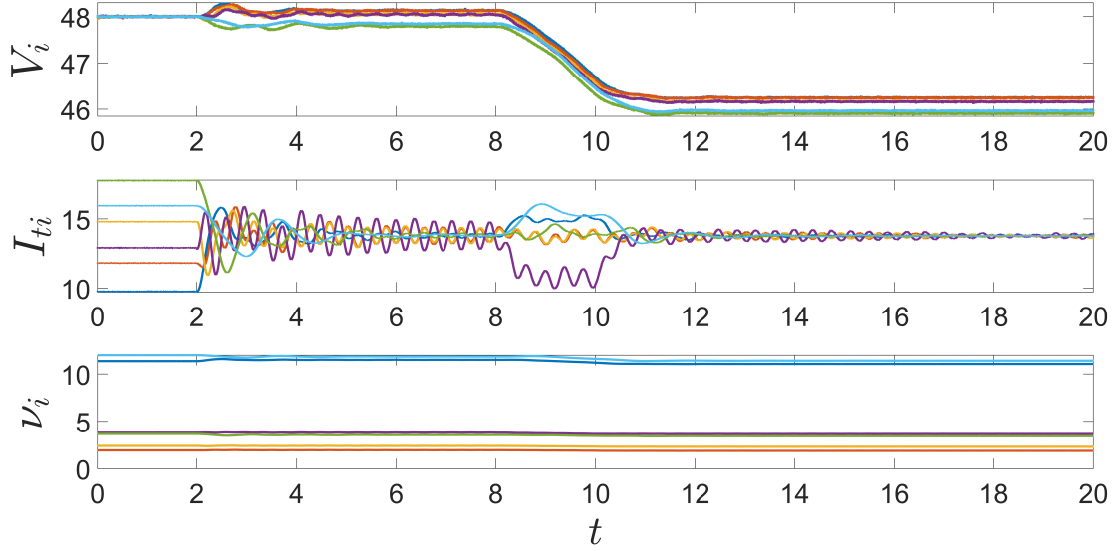


Figure 4.7: Effect of the covert cyber-attacks α_{24} and α_{34} on the DC microgrid as a whole. Plots of each state for all DGUs $i \in \mathcal{N} = \{1, \dots, 6\}$.

4.6 Conclusions

In this chapter, we have presented a distributed monitoring architecture capable of detecting attacks in large-scale systems subject to stochastic uncertainties. We show how an unbiased minimum-variance estimator can be adapted to a decentralized estimation filter, and how the communication between subsystem controllers can be monitored through it. Structural limitations of the architecture are recognised, which further illustrate the results in Chapter 3.2. Thus, additional monitoring units are designed to counteract this.

Finally, simulation results based on the DC microgrid case-study have been shown, demonstrating the effectiveness of the method. In the following chapter, we show how these results can be applied to real world conditions, by presenting experimental results on a network of Electric Springs.

Chapter 5

Distributed attack detection for a network of electric springs: Experimental results

In the previous chapters, we have presented a distributed cyber-attack detection architecture to monitor the information transmitted between controllers, and we have shown its effectiveness through simulations. In this chapter we provide some results which were achieved experimentally, demonstrating how this framework may be applied to a real world scenario. These experiments were carried out at the Power Electronics Laboratory, Department of Electrical and Electronic Engineering, University of Hong Kong, thanks to the generous hospitality of Prof Shu Yuen Ron Hui. We must note that experimental results were carried out on a network of electric springs, rather than a DC microgrid. The reason for this is twofold: on the one hand, practically, it was not possible to fully test the proposed methodology on an experimental DC microgrid setup; on the other hand, by validating our method on a different system, we show how the distributed detection architecture we present can be applied to a wider class of systems than those considered in this thesis.

Electric springs are a demand-side management technology to regulate voltage amplitude and frequency in AC distribution networks. Initially introduced in [HLW12], electric springs are designed to improve the voltage and power quality of distribution networks [STLH14]. Indeed,

with the increasing presence of distributed renewable energy sources within the electrical grid, the need to develop and implement automatic systems to regulate power quality for certain critical nodes has become paramount. Electric springs are a contribution in this direction.

Among recent developments in the distributed control of electric springs, [CYY⁺19] has introduced a consensus-based control architecture, allowing for a network of electric springs to equally contribute to the injected power, thus achieving power sharing. To enable cooperation [CYY⁺19] proposes a consensus-based control architecture, itself requiring the inclusion of a communication network between the cluster of electric springs. Thus, it is an ideal test-bed to verify the distributed cyber-attack detection algorithms developed throughout this work.

In the following, after presenting the model of a single electric spring, we summarise the control protocol in [CYY⁺19], showing how attacks may alter the behaviour of the overall system. Thus, we adapt the architecture presented in Chapter 4, modifying it to account for approximation errors and nonlinearities. Finally, we present the experimental results.

5.1 Networks of Electric Springs

Electric springs are a smart-grid technology, developed to improve the power quality of electrical networks with large penetration of renewable energy sources [HLW12]. Operating as smart loads, they alter the power consumption of non-critical loads such that the power quality at some *critical* nodes are properly regulated. A string of electric springs within an AC microgrid setting is shown in Figure 5.1.

5.1.1 Physical characteristics of the electric spring

Electric springs are composed of a DC voltage source (such as a battery), a DC-AC inverter, here implemented via a half-bridge, and an LC filter, filtering out the high-order harmonic components of the AC waveform from the switching devices. As seen from the circuit diagram in Figure 5.2, the LC filter is then connected in series with a *non-critical load*, i.e. a load that can sustain some degree of voltage and frequency fluctuations without unacceptable drops in

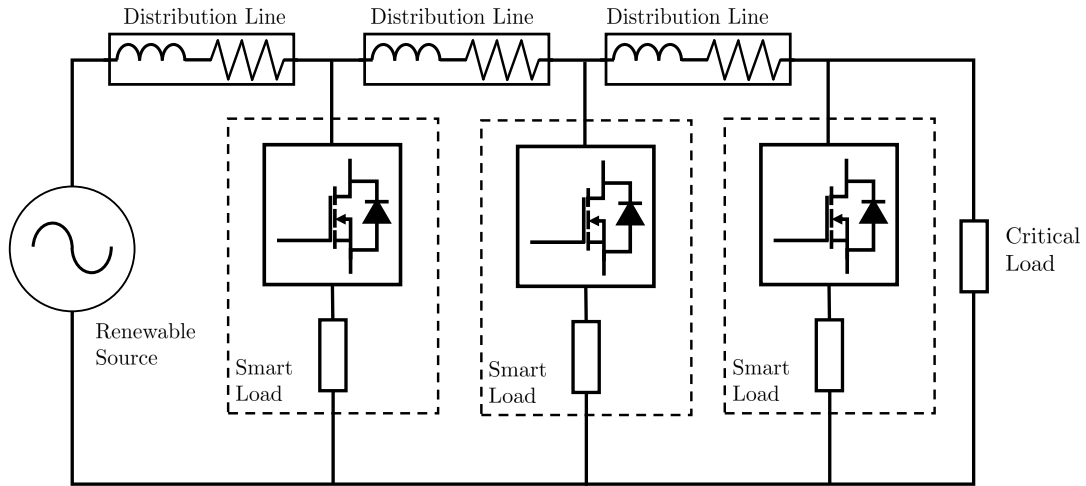


Figure 5.1: Diagram of the physical layer of an AC microgrid with a cluster of electric springs. The electric springs (or smart loads) are connected in parallel with each other. The network of electric springs is set to regulate the supply line voltage and frequency for a so-called “critical” load, for which power regulation is important for proper performance.

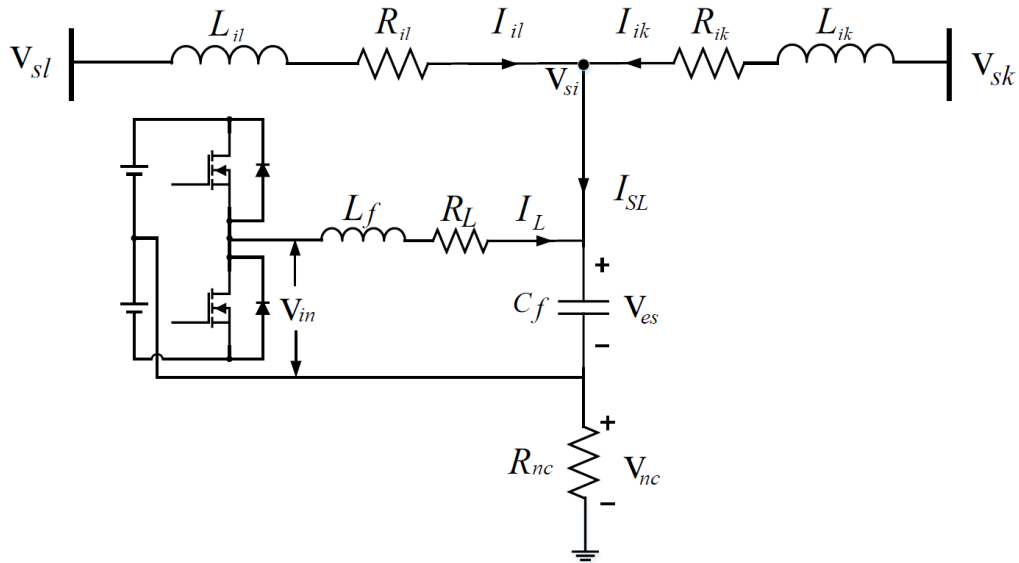


Figure 5.2: Circuit diagram of an electric spring ES_i , with neighbours $\mathcal{N}_i \doteq \{l, k\}$.

performance. This component is crucial for the operations of the electric spring, as it allows the smart load to change its power consumption to regulate the line voltage. Indeed, the electric spring operates by adaptively altering the voltage of the non-critical load, thus changing its power consumption to regulate the line voltage V_s . This is shown in Figure 5.3, using the dq transformation. The dq-transformation, or Park’s transformation [WCJ⁺12], is a method of

representing an AC waveform in a rotating frame as a DC level. We assume the following.

Assumption 5.1. *For the line voltage waveform $V_s(t) = |V_s(t)|\sin(\omega(t)t + \phi(t))$ with nominal frequency ω^{nom} , phase-shift ϕ^{nom} and magnitude V_s^{nom} , we assume that*

$$\begin{cases} \omega(t) = \omega^{\text{nom}}, \\ \phi(t) = \phi^{\text{nom}} \end{cases} \quad (5.1)$$

for all $t \geq 0$, while $|V_s(t)| \neq V_s^{\text{nom}}$. $\triangleleft$

Remark 5.1. In this experimental set-up, as a consequence of Assumption 5.1, we suppose that the smart load must only contribute reactive power to stabilise source voltage V_s . $\triangleleft$

For the purposes of analysis and control, we introduce the dq-decomposition of the AC voltage waveform, which simplifies the analysis of the dynamics and the control throughout this section. Selecting a reference signal, all AC waveforms with the same frequency can then be decomposed into their *direct* and *quadrature* components. This in turn simplifies analysis of AC waveforms significantly.

In the context of this chapter, the non-critical load current is taken as the rotating reference about which all other signals are transformed. In Figure 5.3, we show how the source voltage V_s , the non-critical load voltage V_{nc} , and the electric spring voltage V_{es} are represented in this analytical framework. Note that V_{nc} is in phase with the reference I_{nc} , as we are considering a purely resistive non-critical load R_{nc} , as shown in Figure 5.2. Furthermore, given that the electric spring can be controlled to regulate both components independently [STLH14], the magnitude and phase of V_s can be changed by changing $V_{es}^{d/q}$. We omit further details of the operation of the electric spring, but refer to [HLW12, STLH14] for completeness.

5.1.2 State-space modelling of electric spring dynamics

Considering a set of interconnected electric springs in parallel on a distribution power line, as shown in Figure 5.1, let us present the dynamics of the i -th electric spring ES_i . We make use

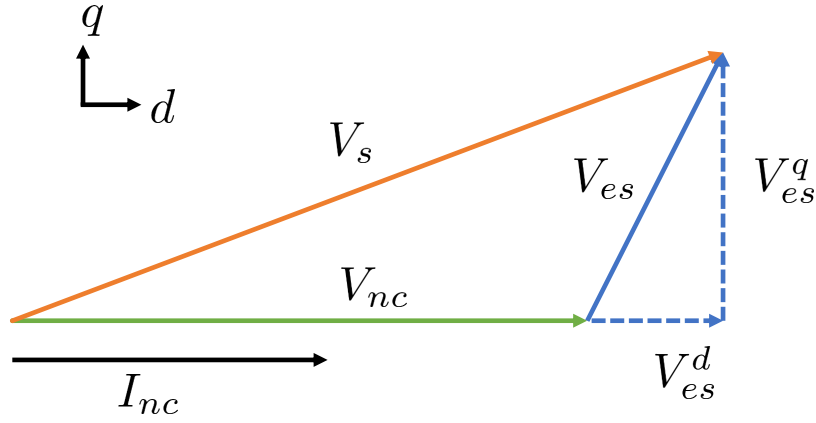


Figure 5.3: Park decomposition (or dq decomposition) of voltages in an electric spring. The dq-transformation, or Park's transformation [WCJ⁺12], is a method of representing an AC waveform in a rotating frame as a DC level. Non-critical load current I_{nc} [black] is taken as the reference for the decomposition. Given that non-critical load is considered to be purely resistive, $V_{nc}^q = 0$, and V_{nc} [green] is aligned with the d axis of the diagram. Electric spring voltage V_{es} [blue] is decomposed into V_{es}^d and V_{es}^q [dotted blue]. The voltage at the point of common coupling V_s [red] is given by Kirchhoff's voltage law. The direct and quadrature axes of the diagram are shown in the top left-hand corner.

of a decoupled single phase dq-transformation, with the non-critical load current I_{nc_i} taken as reference.

The complete dynamics of the ES_i can be defined as the following differential equations¹:

$$\left\{ \begin{array}{l} C_f \dot{V}_{es}^d = \frac{1}{X_{C_f}} V_{es}^q + I_L^d + \sum_{j \in \mathcal{N}_i} I_{ij}^d \\ C_f \dot{V}_{es}^q = \frac{1}{X_{C_f}} V_{es}^d + I_L^q + \sum_{j \in \mathcal{N}_i} I_{ij}^q \\ L_f \dot{I}_L^d = X_{L_f} I_L^q + V_{in}^d - V_{es}^d - R_L I_L^d \\ L_f \dot{I}_L^q = -X_{L_f} I_L^d + V_{in}^q - V_{es}^q - R_L I_L^q \\ L_{ij} \dot{I}_{ij}^d = X_{L_{ij}} I_{ij}^q + V_{s_j}^d - V_{es}^d - R_{ij} I_{ij}^d - R_{nc} I_{nc}^d \\ L_{ij} \dot{I}_{ij}^q = -X_{L_{ij}} I_{ij}^d + V_{s_j}^q - V_{es}^q - R_{ij} I_{ij}^q - R_{nc} I_{nc}^q \end{array} \right. \quad (5.2)$$

where V_{es} is the electric spring's voltage, taken across the filter capacitor C_f , I_L is the inductor current through L_f , I_{ij} is the line current between ES_i and its neighbours ES_j , $j \in \mathcal{N}_i$, where $\mathcal{N}_i$ is the set of neighbouring electric springs of i , defined as $\mathcal{N}_i \doteq \{j \in \mathcal{N} | R_{ij} < \infty\}$, with $\mathcal{N} \doteq \{1, \dots, N\}$ the set of all indexes of the electric springs in the network. Moreover, V_{in}

¹Here, when possible, we omit the index i to improve legibility.

is the voltage across the DC-AC converter, which acts as the input to the filter, and V_{s_k} is the line (or source) voltage of the k -th electric spring at the point of common coupling. The electrical parameters of the RL filter are defined as C_f , L_f , and R_L , denoting respectively the filter's capacitance and inductance, as well as the inductor's series resistance. Further electrical parameters L_{ij} and R_{ij} are the inductance and resistance of the power line connecting ES_i to an electric spring ES_j in its set of neighbours $\mathcal{N}_i$. Finally, X_C and X_L are the capacitance and inductance reactance, i.e.:

$$X_L = \omega L, \quad X_C = \frac{1}{\omega C}, \quad (5.3)$$

where $\omega = 2\pi f$ is the nominal angular frequency of the AC waveforms in [rad/s], and f is the nominal frequency in [Hz]. Specifically, we consider the nominal frequency $f = 50$ Hz.

By taking Kirchhoff's current and voltage laws at the point of common coupling between the smart load and the distribution line, we can algebraically rewrite the non-critical load current as

$$I_{nc_i} = \sum_{j \in \mathcal{N}_i} I_{ij}^d \quad (5.4)$$

and ES_i 's source voltage V_s at the point of common coupling as

$$\begin{cases} V_s^d = V_{es}^d + R_{nc} I_{nc} \\ V_s^q = V_{es}^q \end{cases} \quad (5.5)$$

Note that I_{nc_i} is considered to be the reference of the local Park transformation, and therefore $I_{nc_i} = I_{nc_i}^d$ and $I_{nc_i}^q = 0$. This, however, does not hold for I_{nc_j} , $j \in \mathcal{N}_i$ in the reference frame defined by I_{nc_i} , i.e. for the load currents of the neighbouring electric springs. Indeed, although they are the reference of the dq-frame for ES_j , these references do not in general coincide for ES_i and ES_j , $i \neq j$, i.e. I_{nc_i} and I_{nc_j} are generally out of phase.

Equations (5.2), (5.4) and (5.5) can be rewritten in state space form, characterizing the dynamics of the ES_i as:

$$\begin{cases} \dot{x}_i = A_{ii}x_i + B_i u_i + G_i \xi_i \\ y_i = C_i x_i \end{cases} \quad (5.6)$$

$$\begin{aligned}
A_{ii} &\doteq \begin{bmatrix} 0 & \omega & \frac{1}{C_f} & 0 & \frac{1}{C_f} & 0 \\ -\omega & 0 & 0 & \frac{1}{C_f} & 0 & \frac{1}{C_f} \\ -\frac{1}{L_f} & 0 & -\frac{R_L}{L_f} & \omega & 0 & 0 \\ 0 & -\frac{1}{L_f} & -\omega & -\frac{R_L}{L_f} & 0 & 0 \\ -\frac{1}{L_{ij}} & 0 & 0 & 0 & -\frac{R_{ij}+R_{nc}}{L_{ij}} & \omega \\ 0 & -\frac{1}{L_{ij}} & 0 & 0 & -\omega & -\frac{R_{ij}+R_{nc}}{L_{ij}} \end{bmatrix}, \quad B_i \doteq \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \frac{1}{L_f} & 0 \\ 0 & \frac{1}{L_f} \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \\
G_i &\doteq \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ \frac{1}{L_{ij}} & 0 \\ 0 & \frac{1}{L_{ij}} \end{bmatrix}.
\end{aligned} \tag{5.7}$$

with $x_i \doteq \left[V_{es}^{d/q^\top}, I_L^{d/q^\top}, I_{ij,j \in \mathcal{N}_i}^{d/q^\top} \right]^\top$ the state of ES_i , $u_i = \left[V_{in_i}^{d/q} \right]$ its control input, $\xi_i = \text{col}_{j \in \mathcal{N}_i} \left(V_{s_j}^{d/q} \right)$ the effect of physical coupling with its neighbours, and $y_i \doteq \left[V_{es}^{d/q^\top}, I_L^{d/q^\top}, I_{nc}^{d/q^\top} \right]^\top$ its output. For a voltage or current waveform $\mathcal{J}$, the notation $\mathcal{J}^{d/q}$ defines the column concatenation of its direct and quadrature components. The definition of matrices A_{ii}, B_i, G_i and E_i can be found in (5.7), while C_i is defined by:

$$C_i \doteq \begin{bmatrix} I_4 & 0 \\ 0 & \text{row}_{j \in \mathcal{N}_i}(I_2) \end{bmatrix}. \tag{5.8}$$

Remark 5.2. Note that when defining the matrices in (5.7) of the state-space dynamic system (5.6), we have considered $I_{ij}(t)$ as a single line current, i.e. as if $N_i = |\mathcal{N}_i| = 1$. For $N_i > 1$, the size of the local state increases to $n_i = 4 + 2N_i$, and the definitions of A_{ii} , B_i , G_i , and E_i are obtained from (5.2). $\triangleleft$

Note that ES_i is physically coupled to its neighbours' states through the term $G_i \xi_i$ in (5.6). Indeed, recalling (5.5),

$$\xi_i \doteq \text{col}_{j \in \mathcal{N}_i} (H_j x_j), \tag{5.9}$$

with matrix H_j defined as:

$$H_j \doteq \begin{bmatrix} 1 & 0 & 0 & 0 & R_{nc_j} & 0 \\ 0 & 1 & 0 & 0 & 0 & R_{nc_j} \end{bmatrix}. \quad (5.10)$$

Furthermore, G_i is full column rank by definition, which is necessary to construct the distributed cyber-attack detection strategy in Section 5.2.

5.1.3 Consensus-based distributed control of electric springs

Following the work of [CYY⁺19], we analyse a distributed consensus control protocol for power sharing amongst the network of electric springs. Given Assumption 5.1, here we consider voltage fluctuation in the distribution line. Thus, the electric springs contribute reactive power for stabilization of local source voltage V_{s_i} .

Remark 5.3. Similarly to current sharing in the case of DC microgrids presented in Chapter 2, reactive power sharing can be beneficial for a string of electric springs. Indeed, by contributing equally to the overall power compensation, further objectives may be achieved by each electric spring, such as maintaining similar levels of charge of the DC-side source battery, allowing the inclusion of lower-rated devices in the network, among others. $\triangleleft$

Let us define Q_i^{es} as the reactive power that ES_i contributes to the network for voltage regulation. It is well known that the reactive power is defined by:

$$Q_i^{es} = V_{es_i}^d I_{nc_i}^q - V_{es_i}^q I_{nc_i}^d. \quad (5.11)$$

Therefore, given a separate control loop for $V_{es_i}^{d/q}$, we set the reference for $V_{es_i}^d = 0$ V, for all $i \in \mathcal{N}$, while regulating the q -axis component of the voltage, thus obtaining:

$$Q_i^{es} = -V_{es_i}^q I_{nc_i}^d. \quad (5.12)$$

To achieve global reactive power sharing, each electric spring must communicate with its

neighbours. Specifically, here we consider that each ES_i receives the data:

$$\gamma_{ji}^c(t) = [V_{es_j}(t), I_{nc_j}(t)]^\top + \beta_{ji}(t - T_a^{ij})\alpha_{ji}(t) \quad (5.13)$$

from each of its neighbours $j \in \mathcal{N}_i$ over a communication network, in turn transmitting its own. Thus, the consensus based control layer is exposed to attacks, modelled by $\beta_{ji}(t - T_a^{ij})\alpha_{ji}(t)$, where α_{ji} is the attack signal, and $\beta_{ji}(\cdot)$ is the activation function, as defined in (3.8).

Remark 5.4. Note that here we consider the AC waveforms, rather than their dq-components, to be transmitted over the communication network. This allows us to more easily implement the distributed attack detection architecture. Indeed, I_{nc_j} and V_{es_j} can therefore be decomposed according to the reference waveform I_{nc_i} , thus finding $V_{s_j}^{d/q}$. This, in turn, permits the dynamics of ES_i to be written as in (5.6). $\triangleleft$

Once ES_i receives the communicated information, it decomposes γ_{ji}^c into its dq components using its local dq frame, thus defining $\gamma_{ji}^{c,d/q}$. The reactive power Q_{ji}^{es} can then be easily computed locally². To regulate $V_{es_i}^q$, a local correction term is computed

$$\Delta V_i = k_Q \sum_{j \in \mathcal{N}_i} a_{ij}(Q_{ji}^{es} - Q_i^{es}) \quad (5.14)$$

with coupling parameter k_Q common to all subsystems.

The resulting correction term is then used to modify the reference voltage V_{ref} in the PI control law in the following:

$$u_i^q = K_P^q(V_{\text{ref}} + \Delta V_i - V_{s_i}) + K_I^q \int_0^t V_{\text{ref}} + \Delta V_i(\tau) - V_{s_i}(\tau) d\tau, \quad (5.15)$$

where K_P^q and K_I^q are the q-axis proportional and integral gains, and where $V_{\text{ref}} = V_s^{\text{nom}}$ defined in Assumption 5.1.

A similar PI controller can be designed for the direct component of V_{es_i} :

$$u_i^d = -K_P^d V_{es_i}^d - K_I^d \int_0^t V_{es_i}^d(\tau) d\tau, \quad (5.16)$$

²Note that we retain the index ji in the definition of reactive power Q_{ji}^{es} , to highlight that it is locally computed by ES_i from received, and possibly corrupted, information.

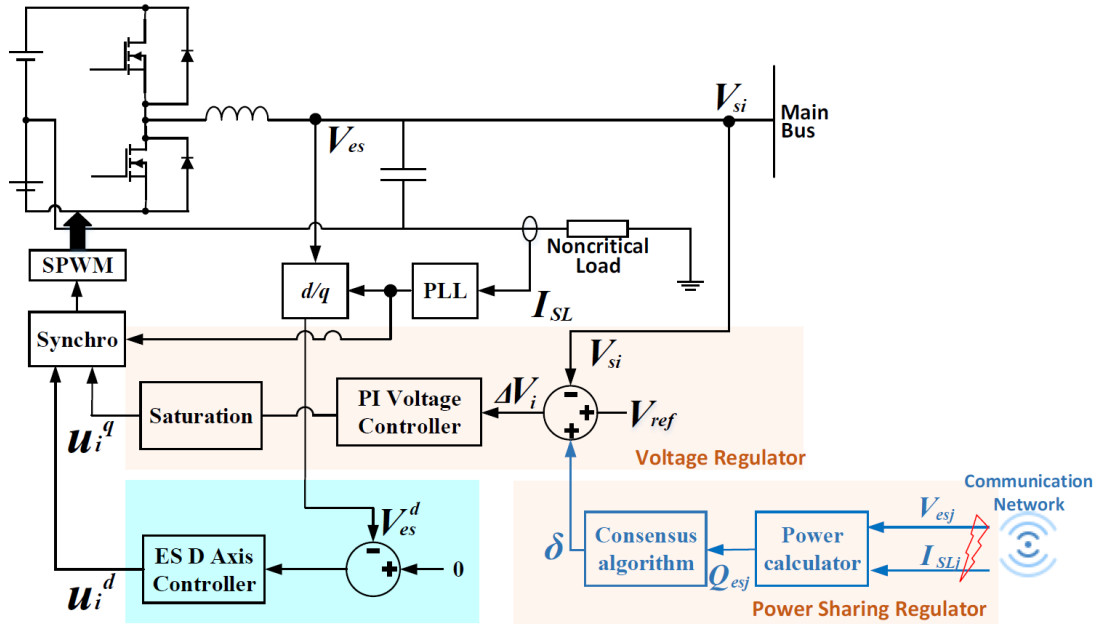


Figure 5.4: Diagram of the distributed controller which is regulating ES_i .

with K_P^d and K_I^d the respective gains, and having set the reference for $V_{es_i}^d$ to 0 V.

To obtain V_{in} from (5.15)-(5.16), the inputs $u_i^{d/q}$ are multiplied by a scaling factor, dependent on the voltage on the DC side of the DC-AC converter in the electric spring, and then to a sinusoidal carrier with the appropriate phase and frequency. This waveform is constructed based on the reference set by I_{nc_i} , which can be found by using a phase lock loop (PLL). The sinusoidal input u_i is then converted into a pulse-width modulated (PWM) signal, used to actuate the switches in the DC-AC converter. The distributed control scheme is shown pictorially in Figure 5.4.

5.2 Distributed cyber-attack detection

5.2.1 Discretisation of electric spring dynamics

Given the experimental nature of the results we present in this chapter, we apply the detection framework presented in Chapter 4, as it allows the minimization of the estimation error variance. Thus, together with the fact that we run experiments using a digital controller, the dynamics of each electric spring are to be discretised.

Similarly to the discretisation method for DC microgrids presented in Chapter 2, we compute the discretisation subsystem-by-subsystem, not to alter the structural relations between them. We must therefore introduce the following assumption.

Assumption 5.2. *The control input $u_i(t)$ and interconnection term $\xi_i(t)$, are considered to be constant for all $t \in [kT_s, (k+1)T_s)$, with T_s the sampling period, for all $k \in \mathbb{N}_0$. $\triangleleft$*

Thus, the discrete time dynamics of ES_i can be written as:

$$\begin{cases} x_i^+ = \bar{A}_{ii}x_i + \bar{B}_i u_i + \bar{G}_i \xi_i + w_i \\ y_i = C_i x_i + v_i \end{cases} \quad (5.17)$$

where x_i^+ is used as a shorthand for $x_i(k+1)$, and $x_i(k) \equiv x_i(kT_s)$, with a slight abuse of notation. The sampling period is given by T_s . The matrices in (5.17) are defined by

$$\bar{A}_{ii} \doteq e^{A_{ii}T_s}, \quad \bar{B}_i = A_{ii}^{-1}(e^{A_{ii}T_s} - I)B_i, \quad \bar{G}_i = A_{ii}^{-1}(e^{A_{ii}T_s} - I)G_i, \quad (5.18)$$

while the process and measurement noise signals are represented as w_i and v_i , defined as realizations of an i.i.d. Gaussian process, as in Assumption 4.3. Let us here repropose Assumption 4.3 for the sake of completeness.

Assumption 5.3. *Process and measurement noise signals w_i and v_i are independent identically distributed white Gaussian processes*

$$w_i(k) \sim \mathcal{N}(0, W_i) \quad v_i(k) \sim \mathcal{N}(0, V_i) \quad \forall k \geq 0, \quad (5.19)$$

with $W_i \succcurlyeq 0$ and $V_i \succcurlyeq 0$ two known covariance matrices, and are mutually uncorrelated for all $k \geq 0$. The initial state $x_i(0) \doteq x_i^0$ is a realization of the normal distribution

$$x_i^0 \sim \mathcal{N}(\bar{x}_i^0, \Pi_i^0), \quad (5.20)$$

i.e. from a distribution which is independent of $w_i(k)$ and $v_i(k)$ for all k , with known mean $\bar{x}_i^0$ and covariance $\Pi_i^0 \succcurlyeq 0$. $\triangleleft$

Remark 5.5. Differently to the case of DC microgrids in Chapter 2, here we allow the discretisation sampling rate to differ from the switching period of the SMPS converter. This is due to a fundamental difference between the DC-DC and DC-AC converters. In fact, for each distributed generation unit in the DC microgrid, without considering the switching nature of the input to the DC-DC converter, we would introduce a non-negligible steady state error. On the other hand, for electric springs, the error integrates to zero over the period of the AC frequency, i.e. every 20 ms at a the nominal frequency $f = 50$ Hz. $\triangleleft$

5.2.2 State estimation for cyber-attack detection

We now introduce the distributed diagnoser used for in this experimental setting, adapting the architecture presented in Chapter 4 to the specific case considered here.

To limit the computational and communication resources required to implement $\mathcal{D}_i$ for each ES_i , we only include a local module estimating $\hat{x}_i$. Indeed, to include modules $\mathcal{O}_{ji}^d$, other than needing additional resources to estimate $x_j, j \in \mathcal{N}_i$, the information in γ_{ji}^c would have to be expanded. Clearly, this change in $\mathcal{D}_i$ comes at the expense of having possible structurally-stealthy attacks against the distributed diagnoser, namely attacks α_{ij} satisfying $\alpha_{ij} \in \ker H_j$. Note that for the previous scenario we considered, that of DC microgrids, this trade-off was not possible. Indeed, in DC microgrids, the variable used to compute the consensus-based secondary control input, the terminal current I_{tj} , lies within the structural null-space of A_{ij} , for any $i \in \mathcal{N}$, without the need of any model knowledge by the attacker. This is critical, as only equipping $\mathcal{D}_i$ with $\mathcal{O}_i^d$ implies that it would not detect any attack altering the consensus equilibrium of the DC microgrid as a whole. On the other hand, for networks of electric springs, given the consensus protocol in (5.14) and the definition of the coupling matrix G_i in (5.7), attacks are structurally stealthy to $\mathcal{D}_i$ if and only if they lie within $\ker H_j$, i.e. the malicious agent must know the physical parameter of ES_j .

The second change made to $\mathcal{O}_i^d$ is the definition of the unknown input d_i . Indeed, here we define $d_i \doteq u_i$, rather than ξ_i . This is done for a few reasons: on the one hand, including u_i in d_i is beneficial, as it permits us to estimate x_i without the introduction of noise given by the switching nature of u_i . Furthermore, by its definition the estimate of the unknown input

$\hat{d}_i$ includes any additional disturbances and modelling errors that may be present. These can include: the discretisation errors; any mismatch in the real electrical parameters and those used in the definition of the estimator; any calibration errors in the sensors, as well as any dynamics that are not modelled by (5.6). Finally, once discretised, $\bar{B}_i$ and $\bar{G}_i$ have partially overlapping column spaces. Thus, it is not possible to perfectly reconstruct both u_i and ξ_i from the estimate $\hat{d}_i$. This introduces further estimation errors in the computation of $\hat{\xi}_i$, to be used for detection. In the following, we show how we overcome this.

Let us rewrite dynamics of ES_i as with respect to the unknown input $d_i = u_i$ as:

$$\begin{cases} x_i^+ = \bar{A}_{ii}x_i + \bar{E}_i d_i + \bar{G}_i \xi_i + w_i \\ y_i = x_i + v_i \end{cases} \quad (5.21)$$

with unknown input vector $d_i^{d/q} \doteq u_i^{d/q}$ and matrix $\bar{E}_i \doteq \bar{B}_i$, where $\bar{B}_i$ is defined in (5.18). Given definition of C_i in (5.7), Assumption 4.4 and Assumption 4.7 hold. Recall that these assumptions were introduced in Chapter 4.2, as they are necessary for the construction of the unbiased minimum variance estimator (4.14).

Further to the definition of $d_i = u_i$, let us introduce a constant output bias vector $\zeta_i \in \mathbb{R}^{n_i}$, to further reduce the influence of any modelling errors on estimate $\hat{x}_i$. We therefore include ζ_i with x_i to form an augmented state

$$\mathbf{x}_i \doteq [x_i^\top, \zeta_i^\top]^\top \quad (5.22)$$

with dynamics

$$\begin{cases} \mathbf{x}_i^+ = \mathbf{A}_{ii}\mathbf{x}_i + \mathbf{G}_i\xi_i + \mathbf{E}_id_i + \mathbf{B}_w w_i \\ y_i = \mathbf{C}_i\mathbf{x}_i + v_i \end{cases} \quad (5.23)$$

with the matrices defined as:

$$\mathbf{A}_{ii} \doteq \begin{bmatrix} \bar{A}_{ii} & 0 \\ 0 & I_{n_i} \end{bmatrix}, \quad \mathbf{G}_i \doteq \begin{bmatrix} \bar{G}_i \\ 0 \end{bmatrix}, \quad \mathbf{E}_i \doteq \begin{bmatrix} \bar{E}_i \\ 0 \end{bmatrix}, \quad \mathbf{B}_w \doteq \begin{bmatrix} I_{n_i} \\ 0 \end{bmatrix}, \quad \mathbf{C}_i \doteq \begin{bmatrix} C_i & I_{n_i} \end{bmatrix}. \quad (5.24)$$

Note that, for matrices defined in (5.24), Assumptions 4.4 and 4.7 hold.

Remark 5.6. There is a trade-off introduced by including the augmented state ζ_i in (5.23). Indeed, given that the variation caused by an attacker introduces a bias in the behaviour of ES_i , part of it is “absorbed” by the estimate $\hat{\zeta}_i$. This in turn reduces the detection capabilities of the diagnoser. However, this trade-off was deemed necessary to improve the steady state performance of the decentralized filter. $\triangleleft$

Having introduced the augmented dynamics (5.23), and defined the unknown input $d_i = u_i$, we can now define the state estimation filter $\hat{x}_i$.

By not having included ξ_i in the unknown input, some estimate $\hat{\xi}_i$ must be included to the dynamics of $\hat{x}_i$, for it not to be biased. We exploit the received communication signal $\gamma_{ji}^c(k)$ to compute the estimate $\hat{\xi}_i(k)$ through the following:

$$\hat{\xi}_i(k) = \text{col}_{j \in \mathcal{N}_i} \left(\tilde{H}_j \gamma_{ji}^c(k) \right) \quad (5.25)$$

with $\tilde{H}_j$ defined as

$$\tilde{H}_j = \begin{bmatrix} 1 & 0 & R_{nc_j} & 0 \\ 0 & 1 & 0 & R_{nc_j} \end{bmatrix}, \quad (5.26)$$

modifying H_j in (5.10) to account for the definition of γ_{ji}^c . Thus, we can define the dynamics of the state estimate as in (4.14):

$$\hat{x}_i(k|k-1) = \mathbf{A}_{ii} \hat{x}_i(k-1|k-1) + \mathbf{G}_i \hat{\xi}_i \quad (5.27a)$$

$$\hat{d}_i(k-1|k) = M_i(k) (y_i(k) - \mathbf{C}_i \hat{x}_i(k|k-1)) \quad (5.27b)$$

$$\hat{x}_i^*(k|k) = \hat{x}_i(k|k-1) + \mathbf{E}_i \hat{d}_i(k-1|k) \quad (5.27c)$$

$$\hat{x}_i(k|k) = \hat{x}_i^*(k|k) + K_i(k) (y_i(k) - \mathbf{C}_i \hat{x}_i^*(k|k)) \quad (5.27d)$$

with matrices $M_i(k), K_i(k)$ defined respectively in (4.15)-(4.16).

To complete the design of the distributed module $\mathcal{O}_i^d$ in $\mathcal{D}_i$, it is necessary to define a detection test. Note that, because $\hat{\xi}_i$ depends on γ_{ji}^c , $\hat{x}_i$ is sensitive to attacks, differently to its

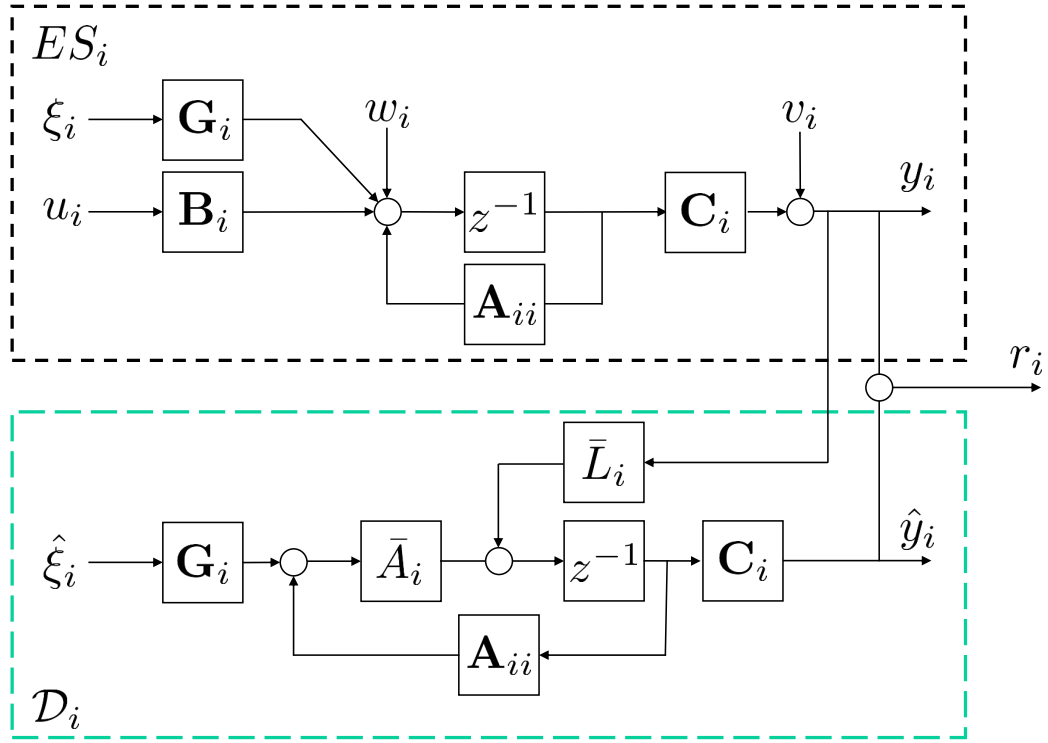


Figure 5.5: Residual computation in $\mathcal{O}_i^d$. To simplify the diagram, we use the compact notation introduced in (4.18), with $\bar{A}_i$ and $\bar{L}_i$ defined in (4.20), rather than (5.27).

equivalent in Chapter 4. It is thus possible to compute the residual as

$$r_i(k) = y_i(k) - \mathbf{C}_i \hat{\mathbf{x}}_i(k), \quad (5.28)$$

and to use it to detect attacks. We present a block diagram representation of the computation of r_i in Figure 5.5.

Let us start by recalling, from the discussion in Chapter 4.2, that the residual $r_i(k)$ can be seen as the output of the LTI system:

$$\begin{cases} \epsilon_i(k) = \bar{A}_i(k) [\mathbf{A}_{ii} \epsilon_i(k-1) + w_i(k-1) + G_i \tilde{v}_i] - \bar{L}_i(k) v_i(k) \\ r_i(k) = \mathbf{C}_i \epsilon_i(k) + v_i(k) \end{cases} \quad (5.29)$$

with $\epsilon_i(k) \doteq \mathbf{x}_i(k) - \hat{\mathbf{x}}_i(k)$ the state estimation error, and where $\tilde{v}_i \doteq \text{col}_{j \in \mathcal{N}_i}(v_j)$, and $\bar{A}_i(k)$, $\bar{L}_i(k)$ are defined in (4.20).

Therefore, given Assumption 5.3 and analysis conducted in Chapter 4, the residual r_i in nominal conditions, i.e. for $k \leq k_a^i$, is a random variable with Gaussian distribution

$$r_i(k) \sim \mathcal{N}(\mu_i^h(k), \Sigma_i(k)), \quad (5.30)$$

with healthy expected value $\mu_i^h(k)$ and covariance matrix $\Sigma_i(k)$. Given the definition of the measurement noise v_i in Assumption 5.3, and the construction of the estimation filter (5.27), the expected value of the residual under healthy conditions is $\mu_i^h(k) = 0$, while the covariance matrix $\Sigma_i(k)$ is defined as:

$$\Sigma_i(k) = \mathbf{C}_i \Pi_i(k) \mathbf{C}_i^\top + V_i - \mathbf{C}_i \bar{L}_i(k) V_i - V_i \bar{L}_i(k)^\top \mathbf{C}_i^\top, \quad (5.31)$$

with $\Pi_i(k) \doteq \text{Var}[\epsilon_i(k)]$, defined by:

$$\Pi_i(k) = \bar{A}_i(k) \left(\mathbf{A}_{ii} \Pi_i(k-1) \mathbf{A}_{ii}^\top + W_i + \mathbf{G}_i \tilde{V}_i \mathbf{G}_i^\top \right) \bar{A}_i(k)^\top + \bar{L}_i(k) V_i \bar{L}_i(k)^\top, \quad (5.32)$$

and where $\tilde{V}_i \doteq \text{diag}_{j \in \mathcal{N}_i} (H_j V_j H_j^\top)$, with H_j defined in (5.10).

Remark 5.7. Note that here $\Pi_i(k)$ depends on $\tilde{V}_j$ rather than Π_j , as we are presuming that $\hat{\xi}_i$ is computed from received measurement signals γ_{ji}^c , rather than estimates $\hat{x}_j$. $\triangleleft$

Recall that $\Sigma_i(k)$ is not invertible [Remark 4.9]. However, because it is symmetric positive-semidefinite, there exists a unitary matrix $U_i(k)$ such that $\Sigma_i(k) = U_i(k) \Lambda_i(k) U_i(k)^\top$. Thus, introducing the auxiliary variable

$$z_i(k) \doteq U_i(k) r_i(k), \quad (5.33)$$

we can implement the following detection test:

$$T'(z_{i,q}, k) = \sum_{\kappa=k-\omega_i+1}^k \frac{z_{i,q}^2(\kappa)}{\lambda_{i,q}(\kappa)} > \vartheta_{i,q}(k) \quad (5.34)$$

where $z_{i,q}$ is the q -th component of z_i , $\lambda_{i,q}$ is the q -th diagonal component of Λ_i , $T'(z_{i,q}, k)$ is a

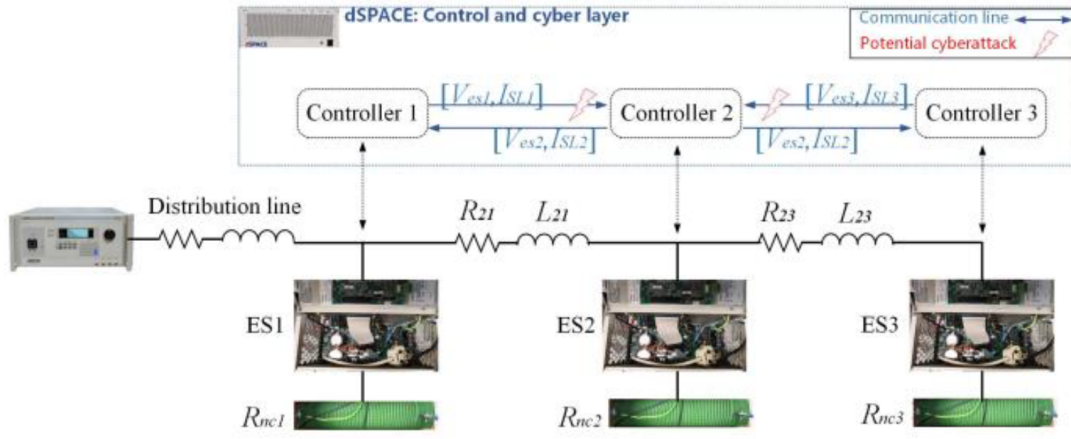


Figure 5.6: Experimental setup of three electric springs in a microgrid. The distribution line is fed by a programmable source with pre-recorder data modelling the effect of renewable energy generation; the three electric springs, each in series with their own non-critical resistive load R_{nc_i} , are in parallel with each other. The control is implemented in a dSPACE digital controller, also implementing the ideal communication between subsystems.

χ^2 detection statistic, and $\vartheta_{i,q}(k)$ is the q -th component of a threshold $\vartheta_i(k) \doteq \text{col}(\vartheta_{i,q})$. The components q are taken such that $\lambda_{i,q} \neq 0$. Further details on the definition and properties of the detection statistic $T'(z_{i,q}(k))$ are given in Chapter 4.3. Moreover, recall from our discussion in Chapter 4.3 that $\vartheta_i(k)$ can be selected to guarantee a suitable false alarm rate $P_i^f(k)$, given the distribution of $T'(z_{i,q}(k))$ in nominal conditions. Having outlined the detection strategy, let us now move on to present the experimental results.

5.3 Experimental results

The hardware setup is set as a 110V AC microgrid. All AC voltages are to be interpreted as root mean square (RMS) values. A California Instruments CSW550 programmable source is used with pre-recorded data to model the effect of renewable energy source injecting time-varying power into the microgrid. We consider a network of $N = 3$ electric springs located along a distribution line, each connected in series with a resistance load, as represented pictorially in Figure 5.6. A picture of the experimental setup in the Power Electronics Laboratory is given in Figure 5.7

The parameters of each electric spring and of the controllers are given in Table 5.1. The control architecture and the ideal communication network are implemented using a dSPACE

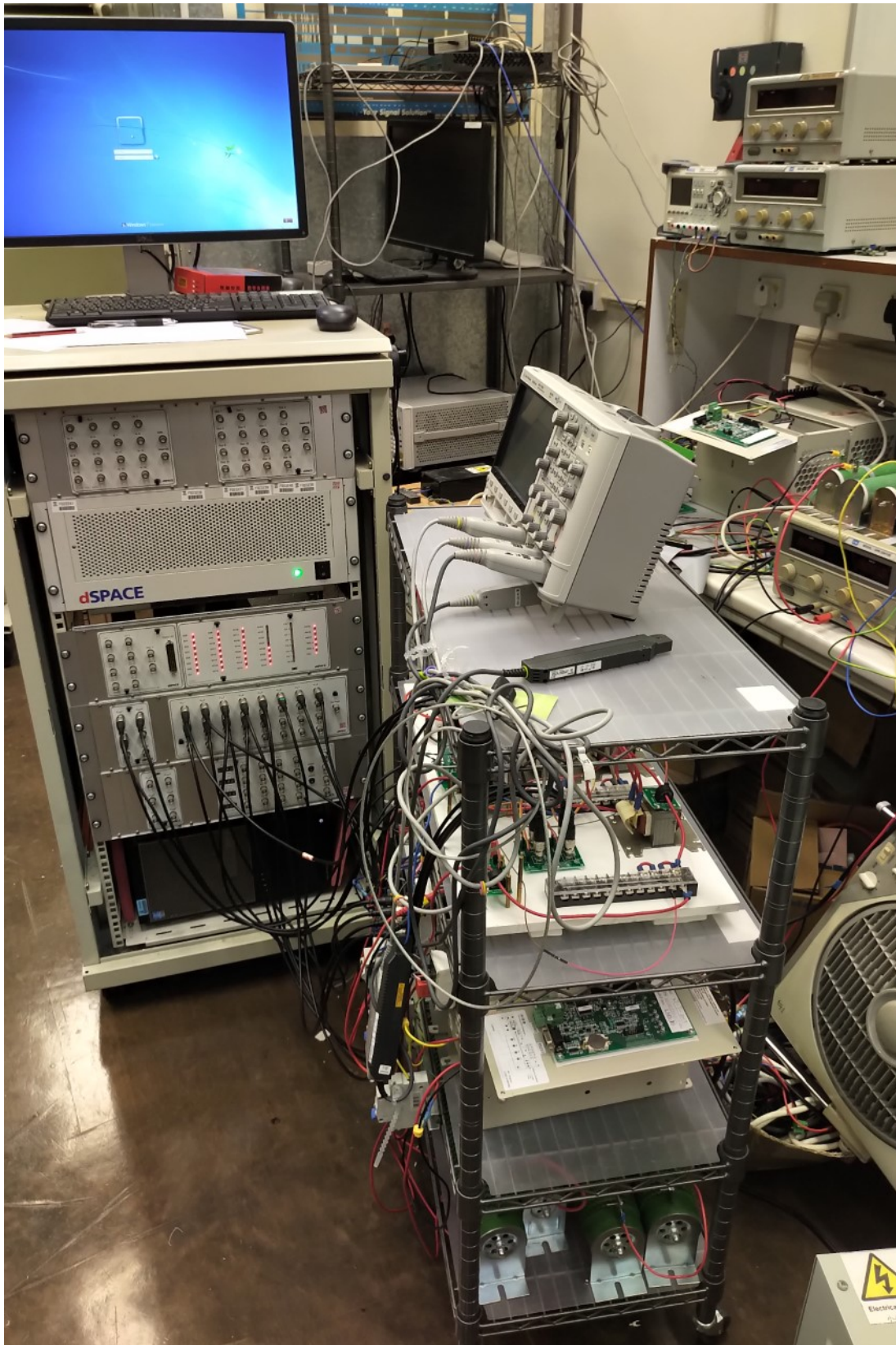


Figure 5.7: A photograph of the experimental setup in the Power Electronics Laboratory at the Department of Electrical and Electronic Engineering, University of Hong Kong.

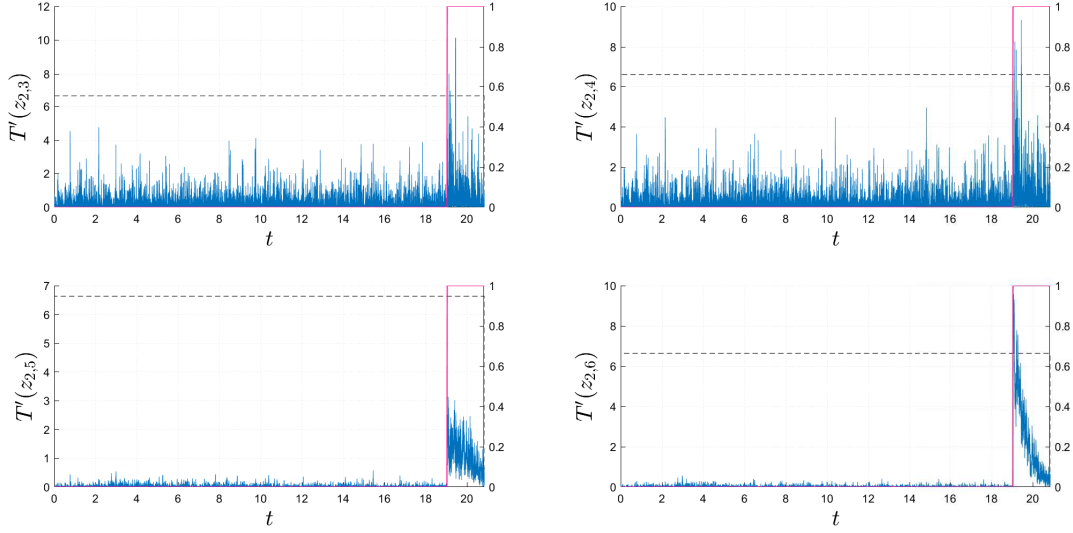


Figure 5.8: Experimental results with bias injection attack of constant amplitude. Statistical tests $T'(z_{2,q})$ for all components q in blue. Threshold $\vartheta_{2,q}$ shown as the dashed black line. In pink, a flag for whether an attack has been detected or not.

controller, collecting sensor measurement data, and interfacing the DC-AC converter through a PWM signal, with sampling frequency $f_s = 10\text{kHz}$. For each ES_i , the noise covariance matrices were defined as

$$\begin{aligned} W_i &\doteq \text{diag}[10^{-2}; 10^{-2}; 10^{-2}; 10^{-2}; 10^{-5}; 10^{-5}; 10^{-5}; 10^{-5}] \\ V_i &\doteq \text{diag}[10^{-2}; 10^{-2}; 10^{-3}; 10^{-3}; 10^{-3}; 10^{-3}]. \end{aligned} \tag{5.35}$$

We consider that communication received by ES_2 from its neighbours in $\mathcal{N}_2 \doteq \{1, 3\}$ to be subject to cyber-attacks, implemented within the communication-layer in the dSPACE controller. Specifically, we consider a bias injection attack on γ_{12}^c , where attack magnitude $\alpha_{12}(k) = 21.21\text{V}$ (RMS) from time $T_a = 19.01\text{s}$ is injected.

In Figure 5.8 we compare the statistical detection test $T'(z_{i,q}, k)$ to the threshold $\vartheta_{i,q}$. The statistical test is computed given its definition in (4.51), while the threshold is selected such that the overall probability of false alarm P_i^f , as defined in (4.57), is $P_i^f = 5\%$. Recalling that the initial time of attack $T_a = 19.01\text{s}$, the detection occurs at time $T_d = 19.04\text{s}$.

5.4 Conclusions

In this chapter, we have provided experimental results demonstrating the effectiveness of the distributed diagnosis algorithm presented in Chapter 4 through experiments. We introduced the electric spring, a device used for demand-side management for voltage balancing in future power networks. Specifically, we gave some insight in its behaviour, and summarized a consensus-based distributed protocol available in literature for reactive power sharing among of networks of electric springs [CYY⁺19]. We then gave indications as to how the framework in Chapter 4 may be explicitly altered in order to better perform on the given hardware. Finally, we presented the experimental setup, and showed our results.

Table 5.1: Experimental setup parameters

Electrical parameters

R_{12}	Line (1, 2) resistance	0.2297Ω
L_{12}	Line (1, 2) inductance	4.428 mH
R_{23}	Line (2, 3) resistance	0.2296Ω
L_{23}	Line (2, 3) inductance	4.426 mH
f_s	Sampling frequency	7.5 kHz
R_{nc1}	Non-critical load ES_1	39.01Ω
R_{nc2}	Non-critical load ES_2	41.58Ω
R_{nc3}	Non-critical load ES_3	40.39Ω
L_{f1}	Filter inductance ES_1	0.5 mH
R_{L1}	Filter resistance ES_1	0.0926Ω
C_{f1}	Filter capacitance ES_1	$6.6 \mu\text{F}$
L_{f2}	Filter inductance ES_2	0.5 mH
R_{L2}	Filter resistance ES_2	0.0926Ω
C_{f2}	Filter capacitance ES_2	$6.6 \mu\text{F}$
L_{f3}	Filter inductance ES_3	0.6 mH
R_{L3}	Filter resistance ES_3	0.1Ω
C_{f3}	Filter capacitance ES_3	$6.65 \mu\text{F}$

Controller parameters

K_P^d	d-axis proportional gain	3
K_I^d	d-axis integral gain	10
K_P^q	q-axis proportional gain	0.3
k_Q	Reactive power consensus coupling gain	0.3

Chapter 6

Conclusion

6.1 Summary of Thesis Achievements

Throughout this thesis, we have presented distributed techniques for detection of cyber-attacks in large-scale cyber-physical systems. Based on the motivating application of DC microgrids controlled via a distributed control architecture for current sharing, we analyse the problem of designing distributed diagnosers with plug-and-play capabilities for a network of physically coupled subsystems that rely on communication for control. We suppose that malicious agents may violate the integrity of this communication network, therefore exposing the system to cyber-attacks.

In Chapter 3, we focus on distributed cyber-attack detection for a class of deterministic subsystems subject to unknown-but-bounded disturbances. Throughout this chapter, we propose a multi-modular distributed architecture, capable of monitoring the information transmitted between controllers for regulation purposes. In Chapter 3.2, we introduce the concept of *reduced-information set*, i.e. a minimum level of information required to detect attacks in large-scale systems, and show that the classes of attacks that can be detected is tightly linked with the physical properties of the coupling between subsystems, by presenting the class of *structurally stealthy* attacks. We then propose a distributed diagnoser $\mathcal{D}_i$ with two parallel and complementary modules, $\mathcal{O}_i^L$ and $\mathcal{O}_{ji}^{UIO}$, both contributing to the overall performance of $\mathcal{D}_i$, but exploiting different *trusted information*. Namely, $\mathcal{O}_i^L$ takes advantage of local trusted measurement outputs, exploiting the physical coupling between subsystems to achieve detection; on the

other hand, $\mathcal{O}_{ji}^{UIO}$ compares the signal received from neighbouring controllers with its expected trajectory. Thus, $\mathcal{D}_i$ is capable of detecting any cyber-attack that does not present the correct dynamical characteristics and that deviates significantly from the actual value. Furthermore, we address the problems of cyber-attack isolation, i.e. the identification of the communication line subject to attacks. Finally, we show through the definition of an algorithmic procedure that this architecture allows for plug-and-play capabilities, so long as the distributed control law does the same.

In Chapter 4, we analyse a similar problem, but in discrete-time and for systems subject to stochastic uncertainties. Based on the same architecture in Chapter 3, we present a decentralised local estimator $\mathcal{O}_i^d$, capable of leveraging the dynamical coupling between subsystems to detect attack. This is coupled with a second module, $\mathcal{O}_{ji}^d$, comparing the received information with its expected trajectory. Note that in this case, the module estimating the local state $\mathcal{O}_i^d$ – corresponding to $\mathcal{O}_i^L$ – is decentralised, and not distributed. This is due to an inherent limitation of partition-based estimation in large-scale interconnected systems. Indeed, without introducing further assumptions, the estimation error covariance depends on the estimation errors of all subsystems of the large-scale system. Conversely, given our decentralised construction, the estimation error depends exclusively on the properties of $\mathcal{S}_i$. Throughout the chapter, we present on the statistical properties of the estimators, and a stochastic detection test used to detect attacks. Both Chapter 3 and Chapter 4 are corroborated through a realistic simulation of a DC microgrid.

Finally, in Chapter 5, we present experimental results that validate the described cyber-attack detection approach. We consider a network of electric springs as the experimental test-bed, controlled by a consensus-based cooperative protocol. Throughout the chapter, we first introduce the dynamic model of the electric spring, a power-electronic device capable of performing demand-side management in electrical distribution networks; we then present the distributed detection architecture, which we adapt from that in Chapter 4 to satisfy some additional design constraints inherent with implementation in a realistic setting. Finally, we present the results of the experiments.

Overall, we have formalised the problem of cyber-attack detection in Chapters 3 and 4 such

that they are well-suited to describing cyber-attacks on the communication network enabling current sharing in DC microgrids. Indeed, we consider attacks only on the information communicated between controllers, rather than supposing the presence of malicious agents between a subsystem and its controller. Furthermore, some of the results presented here, such as the limitations of *reduced-information* detection algorithms presented in Chapter 3.2, are the result of analysis on the considered application.

6.2 Future Work

This thesis, focusing on distributed cyber-attack detection for secure control, can be extended in two natural directions. Namely, these are the inclusion of some non-idealities that were not considered in this thesis, and the design of automatic methods to *accommodate* attacks once they are detected.

Throughout this thesis, a number of assumptions were made, simplifying the analysis of the distributed detection algorithm presented. These include, for both the continuous-time and discrete-time cases, the use of an ideal communication network, and in the discrete-time case, the assumption that the sampling rate for all subsystems was the same and synchronised. As far as the first set of assumptions goes, it is possible to introduce non-idealities in the communication between controllers, such as time delays, which may be asynchronous, and non-deterministic packet delivery. On the other hand, treating the large-scale system as a multi-rate, non-synchronous, and possibly hybrid system would allow for the inclusion of more realistic cases of cyber-physical systems, where the interconnected devices may be greatly heterogeneous. Methods to manage the presence of non-ideal communication networks in control systems are extensively studied in the literature [AB07], and have been tackled from the perspective of distributed fault detection and isolation in [BFK⁺17].

The second major problem that we aim to pursue, control reconfiguration for cyber-attack accommodation, is a fundamental step towards making control systems secure [WS19]. To address this problem, a number of different approaches can be followed.

On the one hand, similarly to the recent work¹ [BP20a], accommodation can be pursued in

¹[BP20a] can be read in its preliminary form in [BP20b].

the direction of estimating the attack to compensate for it. This is a known strategy from fault-tolerant control [BKL⁺06], which is made complex by the specific characteristics of attacks, their coordination and malicious intent. This method for accommodation, however, presents similar hurdles to those faced by the distributed isolation problem tackled in Chapter 3.7. Indeed, depending on the design of the attacks and the structure of the physical coupling between subsystems, it may not be possible to pursue this method for accommodation.

On the other hand, other methods from fault-tolerant control, such as adaptive fault-tolerant control schemes [ZPP04], may be pursued. A promising approach may be one similar to the distributed adaptive scheme presented in [KZC⁺20]. In the paper, an FTC controller based on neural-network adaptive learning techniques is used to achieve consensus for a multi-agent system (MAS) in the presence of additive process faults and multiplicative actuator faults. This technique cannot be used directly, as it considers a network of homogeneous “agents” (subsystems in the vocabulary used in this thesis) without any physical coupling.

Initial efforts in this direction have been made, and we have found some promising preliminary results that must be analysed further.

Bibliography

- [AB07] Panos Antsaklis and John Baillieul. Special issue on technology of networked control systems. Proceedings of the IEEE, 95(1):5–8, 2007.
- [AKP18] Rajasekhar Anguluri, Vaibhav Katewa, and Fabio Pasqualetti. Attack detection in stochastic interconnected systems: Centralized vs decentralized detectors. In 57th IEEE Conference on Decision and Control (CDC), pages 4541–4546, 2018.
- [AKP20] Rajasekhar Anguluri, Vaibhav Katewa, and Fabio Pasqualetti. Centralized versus decentralized detection of attacks in stochastic interconnected systems. IEEE Transactions on Automatic Control, 65(9):3903–3910, 2020.
- [AMP15] F. Arrichiello, A. Marino, and F. Pierri. Observer-based decentralized fault detection and isolation strategy for networked multirobot systems. IEEE Transactions on Control Systems Technology, 23(4):1465–1476, 2015.
- [APKLGH17] Kyriaki E Antoniadou-Plytaria, Iasonas N Kouveliotis-Lysikatos, Pavlos S Georgilakis, and Nikos D Hatziaargyriou. Distributed and decentralized voltage control of smart distribution networks: Models, methods, and future research. IEEE Transactions on Smart Grid, 8(6):2999–3008, 2017.
- [Bak08] Lubomír Bakule. Decentralized control: An overview. Annual Reviews in Control, 32(1):87–98, 2008.
- [Bak14] Lubomír Bakule. Decentralized control: Status and outlook. Annual Reviews in Control, 38(1):71–80, 2014.

- [BCF⁺19] Francesca Boem, Ruggero Carli, Marcello Farina, Giancarlo Ferrari-Trecate, and Thomas Parisini. Distributed fault detection for interconnected large-scale systems: A scalable plug & play approach. IEEE Transactions on Control of Network Systems, 6(2):800–811, 2019.
- [Bej11] Francisco Javier Bejarano. Partial unknown input reconstruction for linear systems. Automatica, 47(8):1751–1756, 2011.
- [BFK⁺17] Francesca Boem, Riccardo M G Ferrari, Christodoulos Keliris, Thomas Parisini, and Marios M Polycarpou. A distributed networked approach for fault detection of large-scale systems. IEEE Transactions on Automatic Control, 62(1):18–33, 2017.
- [BFP09] Francisco J Bejarano, Leonid Fridman, and Alexander Poznyak. Unknown input and state estimation for unobservable systems. SIAM Journal on Control and Optimization, 48(2):1155–1178, 2009.
- [BFP11] Francesca Boem, Riccardo MG Ferrari, and Thomas Parisini. Distributed fault detection and isolation of continuous-time non-linear systems. European Journal of Control, 17(5):603–620, 2011.
- [BG11] Radhakisan Baheti and Helen Gill. Cyber-physical systems. The impact of control technology, 12:161–166, 2011.
- [BGBP19] Angelo Barboni, Alexander J Gallo, Francesca Boem, and Thomas Parisini. A distributed approach for the detection of covert attacks in interconnected systems with stochastic uncertainties. In 2019 IEEE 58th Conference on Decision and Control (CDC), pages 5623–5628. IEEE, 2019.
- [BGFTP17] Francesca Boem, Alexander J Gallo, Giancarlo Ferrari-Trecate, and Thomas Parisini. A distributed attack detection method for multi-agent systems governed by consensus-based control. In 2017 IEEE 56th Annual Conference on Decision and Control (CDC), pages 5961–5966. IEEE, 2017.

- [BGRP20] Francesca Boem, Alexander J Gallo, Davide M Raimondo, and Thomas Parisini. Distributed fault-tolerant control of large-scale systems: an active fault diagnosis approach. IEEE Transactions on Control of Networked Systems, 7(1):288–301, 2020.
- [BKL⁺06] Mogens Blanke, Michel Kinnaert, Jan Lunze, Marcel Staroswiecki, and J Schröder. Diagnosis and fault-tolerant control. Springer, 2006.
- [BKLS16] Mogens Blanke, Michel Kinnaert, Jan Lunze, and Marcel Staroswiecki. Distributed fault diagnosis and fault-tolerant control. In Diagnosis and Fault-Tolerant Control, pages 467–518. Springer, 2016.
- [BL88] Lubomír Bakule and Jan Lunze. Decentralized design of feedback control for large-scale systems. Kybernetika, 24(8):1–3, 1988.
- [BL15] Sven Bodenburg and Jan Lunze. Plug-and-play reconfiguration of locally interconnected systems with limited model information. IFAC-PapersOnLine, 48(22):20–27, 2015.
- [BNJD18] Omar Ali Beg, Luan V Nguyen, Taylor T Johnson, and Ali Davoudi. Signal temporal logic-based attack detection in DC microgrids. IEEE Transactions on Smart Grid, 2018.
- [BP20a] Angelo Barboni and Thomas Parisini. Towards distributed accommodation of covert attacks in interconnected systems. In 2020 IEEE 59th Conference on Decision and Control (CDC). IEEE, 2020.
- [BP20b] Angelo Barboni and Thomas Parisini. Towards distributed accommodation of covert attacks in interconnected systems, 2020.
- [BPG17] Cheng-Zong Bai, Fabio Pasqualetti, and Vijay Gupta. Data-injection attacks in stochastic control systems: Detectability and performance tradeoffs. Automatica, 82:251–260, 2017.

- [BRBP19] Angelo Barboni, Hamed Rezaee, Francesca Boem, and Thomas Parisini. Distributed detection of covert attacks for interconnected systems. In 2019 18th European Control Conference (ECC), pages 2240–2245. IEEE, 2019.
- [BRBP20] Angelo Barboni, Hamed Rezaee, Francesca Boem, and Thomas Parisini. Detection of covert cyber-attacks in interconnected systems: A distributed model-based approach. IEEE Transactions on Automatic Control, 2020.
- [BRFTP15] Francesca Boem, Stefano Rivero, Giancarlo Ferrari-Trecate, and Thomas Parisini. A plug-and-play fault diagnosis approach for large-scale systems. IFAC-PapersOnLine, 48(21):601–606, 2015.
- [BRFTP19] Francesca Boem, Stefano Rivero, Giancarlo Ferrari-Trecate, and Thomas Parisini. Plug-and-play fault detection and isolation for large-scale nonlinear systems with stochastic uncertainties. IEEE Transactions on Automatic Control, 64(1):4–19, 2019.
- [BSS16] F. Boem, L. Sabattini, and C. Secchi. Decentralized fault diagnosis for heterogeneous multi-agent systems. In 2016 3rd Conference on Control and Fault-Tolerant Systems (SysTol), pages 771–776, 2016.
- [BTS13] Jan Bendtsen, Klaus Trangbaek, and Jakob Stoustrup. Plug-and-play control—modifying control systems online. IEEE Transactions on Control Systems Technology, 21(1):79–93, 2013.
- [Bul18] F. Bullo. Lectures on Network Systems. CreateSpace, 1 edition, 2018. With contributions by J. Cortes, F. Dorfler, and S. Martinez.
- [CAS08a] A.A. Cardenas, S. Amin, and S. Sastry. Secure control: Towards survivable cyber-physical systems. In 28th International Conference on Distributed Computing Systems Workshops., pages 495–500. IEEE, 2008.
- [CAS08b] Alvaro A Cárdenas, Saurabh Amin, and Shankar Sastry. Research challenges for the security of control systems. In HotSec, 2008.

- [Cas16] Defense Use Case. Analysis of the cyber attack on the ukrainian power grid. Electricity Information Sharing and Analysis Center (E-ISAC), 2016.
- [CBCZ16] Guido Cavraro, Saverio Bolognani, Ruggero Carli, and Sandro Zampieri. The value of communication in the voltage regulation problem. In 2016 IEEE 55th Conference on Decision and Control (CDC), pages 5781–5786. IEEE, 2016.
- [CCSZ08] Ruggero Carli, Alessandro Chiuso, Luca Schenato, and Sandro Zampieri. Distributed Kalman filtering based on consensus strategies. IEEE Journal on Selected Areas in Communications, 26(4):622–633, 2008.
- [CKFL15] Ilhami Colak, Ersan Kabalci, Gianluca Fulli, and Stavros Lazarou. A survey on the contributions of power electronics to smart grid systems. Renewable and Sustainable Energy Reviews, 47:562–579, 2015.
- [Cla78] Robert N Clark. Instrument fault detection. IEEE Transactions on Aerospace and Electronic Systems, (3):456–465, 1978.
- [CPH⁺20] Zhiyong Chen, Fabio Pasqualetti, Jianping He, Peng Cheng, Harry L. Trentelman, and Francesco Bullo. Guest editorial: Special issue on security and privacy of distributed algorithms and network systems. IEEE Transactions on Automatic Control, 65(9):3725–3727, 2020.
- [CPZ96] Jie Chen, Ron J Patton, and Hong-Yue Zhang. Design of unknown input observers and robust fault detection filters. International Journal of control, 63(1):85–105, 1996.
- [CSS17] Peng Cheng, Ling Shi, and Bruno Sinopoli. Guest editorial special issue on secure control of cyber-physical systems. IEEE Transactions on Control of Network Systems, 4(1):1–3, 2017.
- [CST19] Michelle S Chong, Henrik Sandberg, and André MH Teixeira. A tutorial introduction to security and privacy for cyber-physical systems. In 2019 18th European Control Conference (ECC), pages 968–978. IEEE, 2019.

- [CT98] Martin Corless and Jay Tu. State and input estimation for a class of uncertain systems. Automatica, 34(6):757–764, 1998.
- [CYY⁺19] Jie Chen, Shuo Yan, Tianbo Yang, Siew-Chong Tan, and Shu Yuen Hui. Practical evaluation of droop and consensus control of distributed electric springs for both voltage and frequency regulation in microgrid. IEEE Transactions on Power Electronics, 34(7):6947–6959, 2019.
- [DAAW20] Flavia C. Delicato, Adnan Al-Anbuky, and Kevin I-Kai Wang. Editorial: Smart cyber-physical systems: Toward pervasive intelligence systems. Future Generation Computer Systems, 107:1134–1139, 2020.
- [DHWG19] Derui Ding, Qing-Long Han, Zidong Wang, and Xiaohua Ge. A survey on model-based distributed control and filtering for industrial cyber-physical systems. IEEE Transactions on Industrial Informatics, 15(5):2483–2499, 2019.
- [Din08] Steven X Ding. Model-based fault diagnosis techniques: design schemes, algorithms, and tools. Springer Science & Business Media, 2008.
- [DMK16] Mohammadreza Davoodi, Nader Meskin, and Khashayar Khorasani. Simultaneous fault detection and consensus control design for a network of multi-agent systems. Automatica, 66:185–194, 2016.
- [DPA⁺18] SM Dibaji, M Pirani, AM Annaswamy, KH Johansson, and A Chakraborty. Secure control of wide-area power systems: Confidentiality and integrity threats. In 2018 IEEE Conference on Decision and Control (CDC), pages 7269–7274. IEEE, 2018.
- [DSM⁺16] Pietro Danzi, Cedomir Stefanovic, Lexuan Meng, Josep M Guerrero, and Petar Popovski. On the impact of wireless jamming on the distributed secondary microgrid control. In 2016 IEEE Globecom Workshops (GC Wkshps), pages 1–6. IEEE, 2016.

- [DUSL19] Mohammad Deghat, Valery Ugrinovskii, Iman Shames, and Cédric Langbort. Detection and mitigation of biasing attacks on distributed estimation networks. Automatica, 99:369–381, 2019.
- [DZ97] Mohamed Darouach and Michel Zasadzinski. Unbiased minimum variance estimation for systems with unknown exogenous inputs. Automatica, 33(4):717–719, 1997.
- [FBB⁺80] Wladyslaw Findeisen, Frederic N Bailey, Mieczyslaw Brdys, Krzysztof Malinowski, Piotr Tatjewski, and Adam Wozniak. Control and Coordination in Hierarchical Systems. John Wiley & Sons, 1980.
- [FBC⁺17] Nicola Forti, Giorgio Battistelli, Luigi Chisci, Suqi Li, Bailu Wang, and Bruno Sinopoli. Distributed joint attack detection and secure state estimation. IEEE Transactions on Signal and Information Processing over Networks, 4(1):96–110, 2017.
- [FC16] Marcello Farina and Ruggero Carli. Partition-based distributed Kalman filter with plug and play features. IEEE Transactions on Control of Network Systems, 5(1):560–570, 2016.
- [FIDSDB17] Gabriella Fiore, Alessio Iovine, Elena De Santis, and Maria Domenica Di Benedetto. Secure state estimation for DC microgrids control. In 13th IEEE Conference on Automation Science and Engineering (CASE), pages 1610–1615, 2017.
- [FMC11] Nicolas Falliere, Liam O Murchu, and Eric Chien. W32. stuxnet dossier. White paper, Symantec Corp., Security Response, 5(6):29, 2011.
- [FPP12] Riccardo MG Ferrari, Thomas Parisini, and Marios M Polycarpou. Distributed fault detection and isolation of large-scale discrete-time nonlinear systems: An adaptive approximation approach. IEEE Transactions on Automatic Control, 57(2):275–290, 2012.

- [Fra96] Peter M Frank. Analytical and qualitative model-based fault diagnosis—a survey and some new results. European Journal of control, 2(1):6–28, 1996.
- [FT20] Riccardo MG Ferrari and Andre MH Teixeira. A switching multiplicative watermarking scheme for detection of stealthy cyber-attacks. IEEE Transactions on Automatic Control, 2020.
- [GBMP13] Jayavardhana Gubbi, Rajkumar Buyya, Slaven Marusic, and Marimuthu Palaniswami. Internet of things (iot): A vision, architectural elements, and future directions. Future Generation Computer Systems, 29(7):1645–1660, 2013.
- [GBP20] Alexander J. Gallo, Angelo Barboni, and Thomas Parisini. On detectability of cyber-attacks for large-scale interconnected systems. In 21st IFAC World Congress, pages –, 2020.
- [GCD15] Zhiwei Gao, Carlo Cecati, and Steven X Ding. A survey of fault diagnosis and fault-tolerant techniques—part I: Fault diagnosis with model-based and signal-based approaches. IEEE Transactions on Industrial Electronics, 62(6):3757–3767, 2015.
- [GCLL13] Josep M Guerrero, Mukul Chandorkar, Tzung-Lin Lee, and Poh Chiang Loh. Advanced control architectures for intelligent microgrids – Part I: Decentralized and hierarchical control. IEEE Transactions on Industrial Electronics, 60(4):1254–1262, 2013.
- [GDM07] Steven Gillijns and Bart De Moor. Unbiased minimum-variance input and state estimation for linear discrete-time systems. Automatica, 43(1):111–116, 2007.
- [GF97] E Alcorta Garcia and PM Frank. Deterministic nonlinear observer-based approaches to fault diagnosis: a survey. Control Engineering Practice, 5(5):663–670, 1997.
- [GSC⁺17] Jairo Giraldo, Esha Sarkar, Alvaro A Cardenas, Michail Maniatakos, and Murat Kantarcioglu. Security and privacy in cyber-physical systems: A survey of surveys. IEEE Design & Test, 34(4):7–17, 2017.

- [GTB⁺] Alexander J Gallo, Mustafa S Turan, Francesca Boem, Giancarlo Ferrari-Trecate, and Thomas Parisini. Distributed watermarking for secure control of microgrids under replay attacks. In 7th IFAC Workshop on Distributed Estimation and Control in Networked Systems (NecSys'18), pages 182–187.
- [GTB⁺20] Alexander J Gallo, Mustafa S Turan, Francesca Boem, Thomas Parisini, and Giancarlo Ferrari-Trecate. A distributed cyber-attack detection scheme with application to DC microgrids. IEEE Transactions on Automatic Control, 65(9):3800–3815, 2020.
- [GTN⁺18] Alexander J Gallo, Mustafa S Turan, Pulkit Nahata, Francesca Boem, Thomas Parisini, and Giancarlo Ferrari-Trecate. Distributed cyber-attack detection in the secondary control of DC microgrids. In European Control Conference (ECC), pages 344–349, 2018.
- [GUC⁺18] Jairo Giraldo, David Urbina, Alvaro Cardenas, Junia Valente, Mustafa Faisal, Justin Ruths, Nils Ole Tippenhauer, Henrik Sandberg, and Richard Candell. A survey of physics-based attack detection in cyber-physical systems. ACM Computing Surveys (CSUR), 51(4):1–36, 2018.
- [GVM⁺10] Josep M Guerrero, Juan C Vasquez, José Matas, Luis García De Vicuña, and Miguel Castilla. Hierarchical control of droop-controlled AC and DC microgrids — A general approach toward standardization. IEEE Transactions on Industrial Electronics, 58(1):158–172, 2010.
- [HF⁺18] Kevin E Hemsley, E Fisher, et al. History of industrial control system cyber incidents. Technical report, Idaho National Lab.(INL), Idaho Falls, ID (United States), 2018.
- [HL04] Guang-Da Hu and Mingzhu Liu. The weighted logarithmic matrix norm and bounds of the matrix exponential. Linear algebra and its applications, 390:145–154, 2004.

- [HLW12] Shu Yuen Hui, Chi Kwan Lee, and Felix F Wu. Electric springs—a new smart grid technology. IEEE Transactions on Smart Grid, 3(3):1552–1561, 2012.
- [HP98a] Ming Hou and Ron J Patton. Input observability and input reconstruction. Automatica, 34(6):789–794, 1998.
- [HP98b] Ming Hou and Ron J Patton. Optimal filtering for systems with unknown inputs. IEEE Transactions on Automatic Control, 43(3):445–449, 1998.
- [Hsi00] Chien-Shu Hsieh. Robust two-stage Kalman filters for systems with unknown inputs. IEEE Transactions on Automatic Control, 45(12):2374–2378, 2000.
- [IB97] Rolf Isermann and Peter Ballé. Trends in the application of model-based fault detection and diagnosis of technical processes. Control engineering practice, 5(5):709–719, 1997.
- [Ise06] Rolf Isermann. Fault-Diagnosis Systems: an Introduction from Fault Detection to Fault Tolerance. Springer Science & Business Media, 2006.
- [JPTT14] Karl H Johansson, George J Pappas, Paulo Tabuada, and Claire J Tomlin. Guest editorial special issue on control of cyber-physical systems. IEEE Transactions on Automatic Control, 59(12):3120–3121, 2014.
- [JRRP10] Pertti Järventausta, Sami Repo, Antti Rautiainen, and Jarmo Partanen. Smart grid power system control in distributed generation environment. Annual Reviews in Control, 34(2):277–286, 2010.
- [JY12] Jin Jiang and Xiang Yu. Fault-tolerant control systems: A comparative study between active and passive approaches. Annual Reviews in control, 36(1):60–72, 2012.
- [Kay93] Steven M Kay. Fundamentals of statistical signal processing. Prentice Hall, 1993.
- [Kit87] Peter K Kitanidis. Unbiased minimum-variance linear state estimation. Automatica, 23(6):775–778, 1987.

- [KPU06] C Kambhampati, RJ Patton, and FJ Uppal. Reconfiguration in networked control systems: Fault tolerant control and plug-and-play. IFAC Proceedings Volumes, 39(13):126–131, 2006.
- [KZC⁺20] Mohsen Khalili, Xiaodong Zhang, Yongcan Cao, Marios M Polycarpou, and Thomas Parisini. Distributed fault-tolerant control of multiagent systems: An adaptive learning approach. IEEE Transactions on Neural Networks and Learning Systems, 31(2):420–432, 2020.
- [LAC16] Robert M. Lee, Michael J. Assante, and Tim Conway. Analysis of the cyber attack on the Ukrainian power grid. SANS Industrial Control Systems, 2016.
- [LKZ⁺16] Sergio Lucia, Markus Kögel, Pablo Zometa, Daniel E Quevedo, and Rolf Find-eisen. Predictive control, embedded cyberphysical systems and systems of systems—a perspective. Annual Reviews in Control, 41:193–207, 2016.
- [LLZC19] Lin-Yu Lu, Hao Jan Liu, Hao Zhu, and Chia-Chi Chu. Intrusion detection in distributed frequency control of isolated microgrids. IEEE Transactions on Smart Grid, 10(6):6502–6515, 2019.
- [Lun92] Jan Lunze. Feedback control of large-scale systems. Prentice Hall, 1992.
- [Mar17] Andrea Martinelli. A passivity-based approach to voltage stabilization in dc microgrids. 2017.
- [MAT] Simscape electrical: user’s guide (specialized power systems).
- [MĆ77] Richard D Middlebrook and Slobodan Ćuk. A general unified approach to modeling switching-converter power stages. International Journal of Electronics Theoretical and Experimental, 42(6):521–550, 1977.
- [MDH⁺18] Federico Milano, Florian Dörfler, Gabriela Hug, David J Hill, and Gregor Verbič. Foundations and challenges of low-inertia systems. In Power Systems Computation Conference (PSCC), pages 1–25. IEEE, 2018.

- [MDRP⁺16] Lexuan Meng, Tomislav Dragicevic, Javier Roldán-Pérez, Juan C Vasquez, and Josep M Guerrero. Modeling and sensitivity study of consensus algorithm-based distributed hierarchical control for DC microgrids. IEEE Transactions on Smart Grid, 7(3):1504–1515, 2016.
- [ME10] Anthony R Metke and Randy L Ekl. Security technology for smart grid networks. IEEE Transactions on Smart Grid, 1(1):99–107, 2010.
- [MKB⁺12] Yilin Mo, Tiffany Hyun-Jin Kim, Kenneth Brancik, Dona Dickinson, Heejo Lee, Adrian Perrig, and Bruno Sinopoli. Cyber-physical security of a smart grid infrastructure. Proceedings of the IEEE, 100(1):195–209, 2012.
- [MMM20] Athira M Mohan, Nader Meskin, and Hasan Mehrjerdi. A comprehensive review of the cyber-attacks and cyber-security on load frequency control of power systems. Energies, 13(15):3860, 2020.
- [MMT00] Mihajlo D Mesarovic, Donald Macko, and Yasuhiko Takahara. Theory of Hierarchical, Multilevel, Systems, volume 68. Elsevier, 2000.
- [MST⁺17] Lexuan Meng, Qobad Shafiee, Giancarlo Ferrari Trecate, Houshang Karimi, Deepak Fulwani, Xiaonan Lu, and Josep M Guerrero. Review on control of DC microgrids and multiple microgrid clusters. IEEE Journal of Emerging and Selected Topics in Power Electronics, 5(3):928–948, 2017.
- [Mur07] Richard M Murray. Recent research in cooperative control of multivehicle systems. Journal of Dynamic Systems, Measurement, and Control, 129(5):571–583, 2007.
- [MVW89] Mohammad-Ali Massoumnia, George C Verghese, and Alan S Willsky. Failure detection and identification. IEEE Transactions on Automatic Control, 34(3):316–321, 1989.
- [MWS15] Yilin Mo, Sean Weerakkody, and Bruno Sinopoli. Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs. IEEE Control Systems, 35(1):93–109, 2015.

- [NER09] North American Electric Reliability Corporation NERC. Nerc critical infrastructure protection (CIP) reliability standards, 2009.
- [NI14] Hiroaki Nishino and Hideaki Ishii. Distributed detection of cyber attacks and faults for power systems. IFAC Proceedings Volumes, 47(3):11932–11937, 2014.
- [NNC⁺19] Giuseppe Notarstefano, Ivano Notarnicola, Andrea Camisa, et al. Distributed optimization for smart cyber-physical networks. Foundations and Trends® in Systems and Control, 7(3):253–383, 2019.
- [NST⁺20] Pulkit Nahata, Raffaele Soloperto, Michele Tucci, Andrea Martinelli, and Giancarlo Ferrari-Trecate. A passivity-based approach to voltage stabilization in DC microgrids with zip loads. Automatica, 113:108770, 2020.
- [OS05] Reza Olfati-Saber. Distributed Kalman filter with embedded consensus filters. In 44th IEEE Conference on Decision and Control, and European Control Conference (CDC-ECC’05), pages 8179–8184. IEEE, 2005.
- [OS06] Reza Olfati-Saber. Flocking for multi-agent dynamic systems: Algorithms and theory. IEEE Transactions on automatic control, 51(3):401–420, 2006.
- [OSFM07] Reza Olfati-Saber, J Alex Fax, and Richard M Murray. Consensus and cooperation in networked multi-agent systems. Proceedings of the IEEE, 95(1):215–233, 2007.
- [OSM04] Reza Olfati-Saber and Richard M Murray. Consensus problems in networks of agents with switching topology and time-delays. IEEE Transactions on automatic control, 49(9):1520–1533, 2004.
- [Pat97] Ron J Patton. Fault-tolerant control systems: The 1997 situation,” in ifac symposium on fault detection supervision and safety for technical processes. In IFAC SAFEPROCESS’1997, 1997.

- [PBB12] Fabio Pasqualetti, Antonio Bicchi, and Francesco Bullo. Consensus computation in unreliable networks: A system theoretic approach. IEEE Transactions on Automatic Control, 57(1):90–104, 2012.
- [PDB13] Fabio Pasqualetti, Florian Dörfler, and Francesco Bullo. Attack detection and identification in cyber-physical systems. IEEE Transactions on Automatic Control, 58(11):2715–2729, 2013.
- [PDB15a] Fabio Pasqualetti, Florian Dorfler, and Francesco Bullo. Control-theoretic methods for cyberphysical security: Geometric principles for optimal cross-layer resilient control systems. IEEE Control Systems, 35(1):110–127, 2015.
- [PDB15b] Fabio Pasqualetti, Florian Dörfler, and Francesco Bullo. A divide-and-conquer approach to distributed attack identification. In 2015 IEEE 54th Annual Conference on Decision and Control, pages 5801–5807, 2015.
- [PKC⁺07] Ronald J Patton, Chandrasekhar Kambhampati, Alessandro Casavola, Ping Zhang, Steven Ding, and Dominique Sauter. A generic strategy for fault-tolerance in control systems distributed over a network. European journal of control, 13(2):280–296, 2007.
- [POM⁺18] Kaveh Paridari, Niamh O’Mahony, Alie El-Din Mady, Rohan Chabukswar, Menouer Boubekeur, and Henrik Sandberg. A framework for attack-resilient industrial control systems: Attack detection and controller reconfiguration. Proceedings of the IEEE, 106(1):113–128, 2018.
- [RBA07] Wei Ren, Randal W Beard, and Ella M Atkins. Information consensus in multivehicle cooperative control. IEEE Control Systems, 27(2):71–82, 2007.
- [RBFTP14] Stefano Riverso, Francesca Boem, Giancarlo Ferrari-Trecate, and Thomas Parisini. Fault diagnosis and control-reconfiguration in large-scale systems: a plug-and-play approach. In 53rd IEEE Conference on Decision and Control, pages 4977–4982. IEEE, 2014.

- [RBFTP16] S. Rivero, F. Boem, G. Ferrari-Trecate, and T. Parisini. Plug-and-play fault detection and control-reconfiguration for a class of nonlinear large-scale constrained systems. IEEE Transactions on Automatic Control, 61(12):3963–3978, 2016.
- [RBGP16] Davide M Raimondo, Francesca Boem, Alexander Gallo, and Thomas Parisini. A decentralized fault-tolerant control scheme based on active fault diagnosis. In 2016 IEEE 55th Conference on Decision and Control (CDC), pages 2164–2169. IEEE, 2016.
- [RFFT13] Stefano Rivero, Marcello Farina, and Giancarlo Ferrari-Trecate. Plug-and-play decentralized model predictive control for linear systems. IEEE Transactions on Automatic Control, 58(10):2608–2614, 2013.
- [RFSFT13] Stefano Rivero, Marcello Farina, Riccardo Scattolini, and Giancarlo Ferrari-Trecate. Plug-and-play distributed state estimation for linear systems. In IEEE 52nd Annual Conference on Decision and Control, pages 4889–4894, 2013.
- [SAJ15] Henrik Sandberg, Saurabh Amin, and Karl Henrik Johansson. Cyberphysical security in networked control systems: An introduction to the issue. IEEE Control Systems, 35(1):20–23, 2015.
- [Sca09] Riccardo Scattolini. Architectures for distributed and hierarchical model predictive control—a review. Journal of Process Control, 19(5):723–731, 2009.
- [Ser18] Dimitrios Serpanos. The cyber-physical systems revolution. Computer, 51(3):70–73, 2018.
- [Sha08] Jeff Shamma. Cooperative control of distributed multi-agent systems. John Wiley & Sons, 2008.
- [SHG⁺12] Siddharth Sridhar, Adam Hahn, Manimaran Govindarasu, et al. Cyber-physical system security for the electric power grid. Proceedings of the IEEE, 100(1):210–224, 2012.

- [SID⁺10] S. Stanković, N. Ilić, Ž. Djurović, M. Stanković, and K.H. Johansson. Consensus based overlapping decentralized fault detection and isolation. In 2010 2nd Conference on Control and Fault-Tolerant Systems (SysTol), pages 570–575, 2010.
- [Sil90] Dragoslav D Siljak. Decentralized control of complex systems, volume 184 of Mathematics in science and engineering. Academic Press, 1990.
- [SM07] Jill Slay and Michael Miller. Lessons learned from the maroochy water breach. In International conference on critical infrastructure protection, pages 73–82. Springer, 2007.
- [Smi11] Roy S Smith. A decoupled feedback structure for covertly appropriating networked control systems. IFAC Proceedings Volumes, 44(1):90–95, 2011.
- [Smi15] Roy S Smith. Covert misappropriation of networked control systems: Presenting a feedback structure. IEEE Control Systems, 35(1):82–92, 2015.
- [SMPD18] Subham Sahoo, Sukumar Mishra, Jimmy Chih-Hsien Peng, and Tomislav Dragicevic. A stealth cyber attack detection strategy for DC microgrids. IEEE Transactions on Power Electronics, 2018.
- [Sob19] Blake Sobczak. Denial of Service attack caused grid cyber disruption: DOE. Environment & Energy Publishing, 2019.
- [SP11] Tariq Samad and Thomas Parisini. Systems of systems. The Impact of Control Technology, pages 175–183, 2011.
- [SSK17] Mahdiah S Sadabadi, Qobad Shafiee, and Alireza Karimi. Plug-and-play robust voltage control of dc microgrids. IEEE Transactions on Smart Grid, 9(6):6886–6896, 2017.
- [Sta14] John A Stankovic. Research directions for the internet of things. IEEE Internet of Things Journal, 1(1):3–9, 2014.

- [STLH14] Yan Shuo, Siew-Chong Tan, CK Lee, and SY Ron Hui. Electric spring for power quality improvement. In 2014 IEEE Applied Power Electronics Conference and Exposition-APEC 2014, pages 2140–2147. IEEE, 2014.
- [Sto09] Jakob Stoustrup. Plug & play control: Control technology towards new challenges. In Control Conference (ECC), 2009 European, pages 1668–1683. IEEE, 2009.
- [STSJ11] Iman Shames, André MH Teixeira, Henrik Sandberg, and Karl H Johansson. Distributed fault detection for interconnected second-order systems. Automatica, 47(12):2757–2764, 2011.
- [SVAS78] Nils Sandell, Pravin Varaiya, Michael Athans, and Michael Safonov. Survey of decentralized control methods for large scale systems. IEEE Transactions on Automatic Control, 23(2):108–128, 1978.
- [TMGFT16] Michele Tucci, Lexuan Meng, Josep M Guerrero, and Giancarlo Ferrari-Trecate. Consensus algorithms and plug-and-play control for current sharing in DC microgrids. arXiv preprint arXiv:1603.03624, 2016.
- [TMGFT18] Michele Tucci, Lexuan Meng, Josep M Guerrero, and Giancarlo Ferrari-Trecate. Stable current sharing and voltage balancing in DC microgrids: A consensus-based secondary control layer. Automatica, 95:1–13, 2018.
- [TPSJ12] André Teixeira, Daniel Pérez, Henrik Sandberg, and Karl Henrik Johansson. Attack models and scenarios for networked control systems. In 1st International Conference on High Confidence Networked Systems, pages 55–64. ACM, 2012.
- [TRFT17] Michele Tucci, Stefano Rivero, and Giancarlo Ferrari-Trecate. Line-independent plug-and-play controllers for voltage stabilization in DC microgrids. IEEE Transactions on Control Systems Technology, 26(3):1115–1123, 2017.
- [TRFT18] Michele Tucci, Stefano Rivero, and Giancarlo Ferrari-Trecate. Line-independent plug-and-play controllers for voltage stabilization in DC microgrids. IEEE Transactions on Control Systems Technology, 26(3):1115–1123, 2018.

- [TSH12] Harry L Trentelman, Anton A Stoorvogel, and Malo Hautus. Control theory for linear systems. Springer Science & Business Media, 2012.
- [TSJ10] André Teixeira, Henrik Sandberg, and Karl H Johansson. Networked control systems under cyber attacks with applications to power networks. In American Control Conference, pages 3690–3696, 2010.
- [TSSJ14] André Teixeira, Iman Shames, Henrik Sandberg, and Karl H Johansson. Distributed fault detection and isolation resilient to network model uncertainties. IEEE Transactions on Cybernetics, 44(11):2024–2037, 2014.
- [TSSJ15] André Teixeira, Iman Shames, Henrik Sandberg, and Karl Henrik Johansson. A secure control framework for resource-limited adversaries. Automatica, 51:135–148, 2015.
- [UGC⁺16] David I Urbina, Jairo Giraldo, Alvaro A Cardenas, Junia Valente, Mustafa Faisal, Nils Ole Tippenhauer, Justin Ruths, Richard Candell, and Henrik Sandberg. Survey and new directions for physics-based attack detection in control systems. US Department of Commerce, National Institute of Standards and Technology, 2016.
- [VRK03] Venkat Venkatsubramanian, Raghunathan Rengaswamy, and Surya N Kavuri. A review of process fault detection and diagnosis, part II: Qualitative models and search strategies. Computers & Chemical Engineering, 27(3):313–326, 2003.
- [VRKY03] Venkat Venkatasubramanian, Raghunathan Rengaswamy, Surya N Kavuri, and Kewen Yin. A review of process fault detection and diagnosis: Part III: Process history based methods. Computers & Chemical Engineering, 27(3):327–346, 2003.
- [VRYK03] Venkat Venkatasubramanian, Raghunathan Rengaswamy, Kewen Yin, and Surya N Kavuri. A review of process fault detection and diagnosis: Part i: Quantitative model-based methods. Computers & Chemical Engineering, 27(3):293–311, 2003.

- [vSvN13] Rossouw von Solms and Johan van Niekerk. From information security to cyber security. Computers & Security, 38:97–102, 2013.
- [WCJ⁺12] Birron Mathew Weedy, Brian John Cory, Nick Jenkins, Janaka B Ekanayake, and Goran Strbac. Electric power systems. John Wiley & Sons, 2012.
- [Wil76] Alan S Willsky. A survey of design methods for failure detection in dynamic systems. Automatica, 12(6):601–611, 1976.
- [WL13] Wenye Wang and Zhuo Lu. Cyber security in the smart grid: Survey and challenges. Computer Networks, 57(5):1344–1371, 2013.
- [WM17] Lili Wang and A Stephen Morse. A distributed observer for a time-invariant linear system. IEEE Transactions on Automatic Control, 63(7):2123–2130, 2017.
- [WOM⁺19] Sean Weerakkody, Omur Ozel, Yilin Mo, Bruno Sinopoli, et al. Resilient control in cyber-physical systems: Countering uncertainty, constraints, and adversarial behavior. Foundations and Trends® in Systems and Control, 7(1-2):1–252, 2019.
- [WS15] Sean Weerakkody and Bruno Sinopoli. Detecting integrity attacks on control systems using a moving target approach. In 2015 54th IEEE Conference on Decision and Control (CDC), pages 5820–5826. IEEE, 2015.
- [WS19] Sean Weerakkody and Bruno Sinopoli. Challenges and opportunities: Cyber-physical security in the smart grid. In Smart Grid Control, pages 257–273. Springer, 2019.
- [Yan06] Tai C Yang. Networked control system: a brief survey. IEE Proceedings Control Theory and Applications, 153(4):403, 2006.
- [YBG11] Qiang Yang, Javier A Barria, and Tim C Green. Communication infrastructures for distributed control of power distribution networks. IEEE Transactions on Industrial Informatics, 7(2):316–327, 2011.
- [Zan17] Stefano Zanero. Cyber-physical systems. Computer, 50(4):14–16, 2017.

- [ZD15] Jinxin Zhao and Florian Dörfler. Distributed control and optimization in DC microgrids. Automatica, 61:18–26, 2015.
- [ZMQ⁺18] Heng Zhang, Wenchao Meng, Junjian Qi, Xiaoyu Wang, and Wei Xing Zheng. Distributed load sharing under false data injection attack in an inverter-based microgrid. IEEE Transactions on Industrial Electronics, 66(2):1543–1551, 2018.
- [ZPP02] Xiaodong Zhang, Marios M Polycarpou, and Thomas Parisini. A robust detection and isolation scheme for abrupt and incipient faults in nonlinear systems. IEEE Transactions on Automatic Control, 47(4):576–593, 2002.
- [ZPP04] Xiaodong Zhang, Thomas Parisini, and Marios M Polycarpou. Adaptive fault-tolerant control of nonlinear uncertain systems: an information-based diagnostic approach. IEEE Transactions on Automatic Control, 49(8):1259–1274, 2004.
- [ZYBV15] Xingsi Zhong, Lu Yu, Richard Brooks, and Ganesh Kumar Venayagamoorthy. Cyber security in smart DC microgrid operations. In First IEEE International Conference on DC Microgrids (ICDCM), pages 86–91, 2015.

Appendix A

State estimation in the presence of unknown inputs

The problem of estimating the state of a dynamical system in the presence of unknown inputs has been greatly addressed in various contexts. Indeed, a number of works have addressed the problem of unknown input observers, both in the deterministic and stochastic case. In this Appendix, without any pretence of it being an exhaustive account, we present major results on the following topics:

- a. unknown input observability and reconstruction;
- b. the design of unknown-input observers;
- c. the design of an unbiased minimum-variance estimator.

Regarding all of the topics above, we present fundamental results available in literature: specifically, we rely on [HP98a, Bej11] for what concerns topic unknown input observability and reconstruction; [CPZ96] regarding design and properties of unknown-input observers; and [GDM07] for the minimum-variance unbiased estimator. When talking about unknown-input observers, we also propose a result presented in [GTB⁺20], highlighting some of the limitations of this class of observers, while also presenting a method for solving it.

A.1 Unknown inputs: Observability and reconstruction

We start by briefly presenting some fundamental properties of unknown inputs, namely their observability from the output of the system and their reconstruction. To do so, we rely mainly on [HP98a]. Other foundational papers dealing with unknown input reconstruction exist; specifically, it is worth mentioning [BFP09], which exploits geometric arguments to solve the problem of partial unknown input reconstruction through partial left invertibility of the system dynamics.

Consider the LTI system

$$\begin{cases} \dot{x} = Ax + Bu \\ y = Cx + Du \end{cases} \quad (\text{A.1})$$

where $x \in \mathbb{R}^n$ is the state of the system, $y \in \mathbb{R}^p$ the output, and $u \in \mathbb{R}^q$ the input, which is unknown. We repeat the definition of *input observability* in [HP98a]:

Definition A.1 (Input observability). *An input $u(t)$ is said to be observable if $y(t) = 0$ for $t \geq 0$ implies $u(t) = 0$, $t > 0$.* $\triangleleft$

The input observability problem is therefore to find conditions on the system (A.1) determining whether inputs are input observable. On the other hand, the problem of input reconstruction can be seen as that of defining an estimator capable of estimating it [HP98a, BFP09].

In order to introduce necessary and sufficient conditions for input observability in [HP98a], let us introduce, for a matrix pencil $\lambda M - N$, the set of its eigenvalues

$$\sigma(\lambda M - N) \doteq \{ \bar{\lambda} \in \mathbb{C} \mid \text{rank}(\bar{\lambda} M - N) < \text{normrank}(\lambda M - N) \},$$

where $\text{normrank}(\lambda M - N)$ is the normal (or maximal) rank of the matrix pencil.

Theorem A.1 ([HP98a, Th. 1]). System (A.1) is input observable if and only if:

$$\sigma \begin{bmatrix} \lambda I - A & B \\ C & D \end{bmatrix} = \sigma \begin{bmatrix} \lambda I - A \\ C \end{bmatrix} \quad (\text{A.2})$$

□

[BFP09] offers similar conditions for unknown input observability from a geometric point of view, relying on concepts of left invertibility and controlled and conditioned invariant subspaces. Namely:

Theorem A.2 ([BFP09, Th.4.7]). The following claims are equivalent:

(i) For any initial condition:

$$y(t) = 0, \forall t \geq 0 \implies u(t) = 0, \forall t \geq 0;$$

(ii) The set of invariant zeros of (A.1) is identical to the set of (A, C) -unobservable eigenvalues;

(iii) system (A.1) is left-invertible and $\mathcal{V}^* \equiv \mathcal{N}^*$. □

In the theorem, $\mathcal{V}^*$ is the *weakly unobservable subspace*, i.e. the maximal null-output invariant subspace such that $Ax_0 + Bu \in \mathcal{V}^*$ and $Cx_0 + Du = 0$, for all $x_0 \in \mathcal{V}^*$, and $\mathcal{N}^*$ the (A, C) unobservable-subspace, i.e. $\mathcal{N}^* = \ker \mathcal{O}$, where $\mathcal{O}$ is the (A, C) -observability matrix. Note that the first condition is the definition of observable inputs, while the second is equivalent to condition (A.2).

The condition in [HP98a] for input reconstructability follows (an equivalent condition is stated in [BFP09, Th. 4.10]):

Theorem A.3 ([HP98a, Th. 4]). There exists an estimator for input reconstruction if and only if (A.1) is input observable. □

It has been shown that, in some conditions, in order to reconstruct the unknown input acting upon the system (A.1) it is necessary to use up to r derivatives of the output signal $y(t)$. In [BFP09], the following necessary and sufficient condition on the system dynamics is given to show when this is not necessary:

Theorem A.4 ([BFP09, Th. 4.11]). The vector u can be estimated from the system output y , without using any derivatives, if and only if the following two conditions are fulfilled:

(i) the set of invariant zeros of (A.1) that do not belong to the set of (A, C) -unobservable eigenvalues is in $\mathbb{C}^-$;

(ii)

$$\text{rank} \begin{bmatrix} CB & D \\ D & 0 \end{bmatrix} = q + \text{rank } D.$$

□

Having presented the conditions for which unknown inputs are observable and reconstructable, we now proceed to presenting both unknown-input observers and unbiased minimum-variance filters in the following sections.

A.2 Unknown-input observers

Unknown-input observers (UIOs), as defined in [CPZ96], are linear-time invariant observers capable of generating an asymptotic estimate of the state of a LTI system in the presence of unknown inputs. Specifically, in [CPZ96] a filter robust to unknown disturbances is presented, for the purpose of fault detection and isolation. The design of observers robust to unknown inputs has been tackled by numerous works in literature, from a large number of different points of view. We here do not present an overview of this literature, a summary of which may be found in [Din08].

In the remainder of this section, we present the main results in [CPZ96], as they are of fundamental importance for the results presented in Chapter 3. Furthermore, we analyse the dynamics of the estimation error when the observed LTI system is subject to unknown-but-bounded uncertainties. Finally, we provide some structural limitations on the fault/attack detection properties of UIOs, and give an propose an appropriate method to obtain fault/attack detectability.

A.2.1 Design of UIO

Consider the linear time invariant system with dynamics:

$$\begin{cases} \dot{x} = Ax + Bu + Ed \\ y = Cx \end{cases} \quad (\text{A.3})$$

with $x \in \mathbb{R}^n$, $y \in \mathbb{R}^p$, $u \in \mathbb{R}^m$ the system state, output, and input signals, respectively. Let the matrices A, B, C be known and of appropriate dimensions. The system is exposed to some unknown disturbances, which are modelled through the term Ed , $d \in \mathbb{R}^q$. Matrix $E \in \mathbb{R}^{n \times q}$ is defined such that it is full column rank, i.e. it satisfies

$$\text{rank } E = q, \quad (\text{A.4})$$

without loss of generality. Indeed, as pointed out in [CPZ96], if E is not full column rank, the following decomposition can be implemented

$$Ed = E_1 E_2 d, \quad (\text{A.5})$$

with E_1 full column rank. Note that, as highlighted specifically in [CPZ96], the term Ed can be exploited to model a number of uncertainties, not only additive signals.

Let us offer the definition given in [CPZ96] for *unknown-input observers*.

Definition A.2 (Unknown-input observer [CPZ96, Def.1]). *An observer is defined as an unknown input observer of a LTI system as that in (A.3), if its state estimation error vector e approaches zero asymptotically, regardless of the presence of unknown inputs (disturbances) in the system.* $\triangleleft$

Let us now construct the UIO proposed in [CPZ96]. Define a full-order observer with the following dynamics:

$$\begin{cases} \dot{z} = Fz + TBu + Ky \\ \hat{x} = z + Hy \end{cases} \quad (\text{A.6})$$

where $\hat{x} \in \mathbb{R}^n$ is the state estimate and $z \in \mathbb{R}^n$ is the internal state of the observer. Define matrices F, T, K, H such that the following relations are satisfied:

$$(HC - I)E = 0 \quad (\text{A.7})$$

$$S = I - HC \quad (\text{A.8})$$

$$F = SA - K_1 C \quad (\text{A.9})$$

$$K_2 = FH \quad (\text{A.10})$$

$$K = K_1 + K_2 \quad (\text{A.11})$$

Defining the state estimation error $e = x - \hat{x}$, its dynamics can be derived as:

$$\begin{aligned} \dot{e} &= \dot{x} - \dot{z} - H\dot{y} \\ &= (I - HC)(Ax + Bu + Ed) - Fz - TBu - Ky \\ &= SAx - (HC - I)Ed - Fz - Ky \\ &= (SA - K_1C)x - Fz - K_2y \\ &= F(x - \hat{x}) = Fe \end{aligned} \quad (\text{A.12})$$

which, if F is a Hurwitz-stable matrix, asymptotically approaches zero as $t \rightarrow \infty$. Therefore, (A.6) is a UIO for system (A.3).

The following Theorem offers the necessary and sufficient conditions that must be satisfied by the dynamics (A.3) for the design of the UIO as in (A.6) to be possible:

Theorem A.5 ([CPZ96, Th.1]). Necessary and sufficient for (A.6) to be a UIO for the system defined by (A.3) are:

- i. $\text{rank } CE = \text{rank } E$;
- ii. (C, A_1) is a detectable pair, where

$$A_1 \doteq A - E [(CE)^\top CE]^{-1} (CE)^\top CA.$$

□

It is worth noting [CPZ96] that condition ii. is equivalent to strong observability [TSH12] of the system described by the tuple (A, E, C) , i.e. that

$$\text{rank} \begin{bmatrix} sI - A & E \\ C & 0 \end{bmatrix} = n + q \quad \forall s \in \mathbb{C}_0^+. \quad (\text{A.13})$$

If the conditions in Theorem A.5 hold, the following procedure can be implemented to compute the matrices of (A.6):

- a) Compute H such that $(I - HC)E = 0$;
- b) Compute K_1 such that F is Hurwitz stable;
- c) Compute all other matrices in (A.7)-(A.11).

As pointed out in [CPZ96], K_1 can be seen as the free parameter in the design of a UIO, and indeed the authors exploit this freedom for fault diagnosis purposes. Furthermore, a general solution for matrix H is given as

$$H = E(CE)^\dagger + H_0(I - CE(CE)^\dagger), \quad (\text{A.14})$$

a special solution of which is

$$H^* \doteq E[(CE)^\top CE]^{-1}(CE)^\top. \quad (\text{A.15})$$

Let us now analyse the properties of the estimation error $e(t)$ in the presence of unknown-but-bounded uncertainties.

A.2.2 UIO in the presence of bounded uncertainties

Suppose now that the LTI system (A.3) is subject to some unknown-but-bounded uncertainties, i.e. its dynamics are

$$\begin{cases} \dot{x} = Ax + Bu + Ed + w \\ y = Cx + v \end{cases} \quad (\text{A.16})$$

where $w \in \mathbb{R}^n$ and $v \in \mathbb{R}^p$ are unknown and satisfy

$$|w(t)| \leq \bar{w}, \quad |v(t)| \leq \bar{v}, \quad \forall t \geq 0. \quad (\text{A.17})$$

Taking a UIO filter as defined in (A.6) with matrices satisfying (A.7)-(A.11), we can compute the dynamics of the estimation error and the residual $r \doteq y - C\hat{x}$ to account for these unknown signals:

$$\begin{cases} \dot{e} = Fe + Sw - K_1v - H\dot{v} \\ r = Ce + v \end{cases} \quad (\text{A.18})$$

To remove the dependency of the solution of e on $\dot{v}$, of which nothing is known, we exploit integration by parts. Specifically, the explicit solution of $e(t)$ can be written as:

$$\begin{aligned} e(t) &= e^{Ft}e(0) + \int_0^t e^{F(t-\tau)} [Sw(\tau) - K_1v(\tau) - H\dot{v}(\tau)] d\tau \\ &= e^{Ft} [e(0) + Hv(0)] - Hv(t) + \int_0^t e^{F(t-\tau)} [Sw(\tau) - Kv(\tau)] d\tau. \end{aligned} \quad (\text{A.19})$$

A.2.3 Structural limitations of detectability of UIOs

In this section, we provide results on some structural limitations of UIOs depending on dynamics (A.3), which were presented in [GTB⁺20].

Consider once again the LTI system without unknown disturbances as in (A.3). Let us however omit the control input: indeed, it can be seen from the estimation error dynamics (A.12) that the presence of a known control input does not affect the observer's performance.

Consider, furthermore, that the system is subject to some unknown fault functions f :

$$\begin{cases} \dot{x} = Ax + Ed + B_f f \\ y = Cx + D_f f \end{cases} \quad (\text{A.20})$$

which may influence both the process or the measurements. For the purpose of this Section, let us suppose that

$$B_f = \begin{bmatrix} I_n & 0 \end{bmatrix}, \quad D_f = \begin{bmatrix} 0 & I_p \end{bmatrix}, \quad (\text{A.21})$$

and that $f = [f_n^\top, f_p^\top]^\top$. In the presence of faults, the estimation error and the residual, e and

r respectively, can be derived as:

$$\begin{cases} \dot{e} = Fe + Sf_n - K_1 f_p - H \dot{f}_p \\ r = Ce + f_p \end{cases} \quad (\text{A.22})$$

Thus, if for $f_n \neq 0$ or $f_p \neq 0$, $r \not\rightarrow 0$ as $t \rightarrow \infty$, UIO detects the faults.

In the following Lemma presented in [GTB⁺20] we show how, if $\text{rank } C = \text{rank } E$, the UIO as defined in (A.6) is not suitable for detection of any faults.

Lemma A.1 ([GTB⁺20, Lem. 1]). Suppose a LTI system (A.20) is equipped with UIO (A.6). If the matrices C and E are such that

$$\text{rank } E = \text{rank } C \quad (\text{A.23})$$

then $r \doteq y - C\hat{x}$ is not sensitive to any fault $f \neq 0$, i.e. $r \rightarrow 0$ as $t \rightarrow \infty$, irrespective of f . $\square$

Proof. Exploiting [CT98, Lemma 1], condition i. in Theorem A.5 implies that there exist nonsingular matrices P and Q such that

$$P^{-1}E = \begin{bmatrix} \bar{E}_1 \\ \mathbf{0} \end{bmatrix}, \quad Q^{-1}CP = \begin{bmatrix} \bar{C}_1 & \mathbf{0} \end{bmatrix}, \quad (\text{A.24})$$

where $\bar{E}_1$ and $\bar{C}_1$ satisfy the following:

- a) $\bar{E}_1$ and $\bar{C}_1$ have the same number of rows, equal to $p = q$;
- b) $\bar{E}_1$ is full row rank – and therefore is invertible (given $\text{rank } E = q$);
- c) $\bar{C}_1$ is invertible.

It is possible to construct a UIO for the transformed dynamics for state $\bar{x} = P^{-1}x$ and output $\bar{y} = Q^{-1}y$, noting that conditions i. and ii. hold for the transformed dynamics:

$$\begin{cases} \dot{\bar{x}} = \bar{A}\bar{x} + \bar{E}d \\ \bar{y} = \bar{C}\bar{x} \end{cases} \quad (\text{A.25})$$

having defined

$$\bar{A} \doteq P^{-1}AP, \quad \bar{E} \doteq P^{-1}E, \quad \bar{C} \doteq Q^{-1}CP, \quad (\text{A.26})$$

with $\bar{A}$ a 2-by-2 block matrix

$$\bar{A} = \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \\ \bar{A}_{21} & \bar{A}_{22} \end{bmatrix}.$$

From (A.7) one derives the following matrices, where a bar has been added to highlight their dependence on the transformed system:

$$\begin{aligned} \bar{H} &= \begin{bmatrix} \bar{C}_1^{-1} \\ 0 \end{bmatrix}, \quad \bar{S} = \begin{bmatrix} 0 & 0 \\ 0 & I_{n-q} \end{bmatrix}, \quad \bar{F} = \begin{bmatrix} -\bar{K}_{1,a}\bar{C}_1 & 0 \\ \bar{A}_{21} - \bar{K}_{1,b}\bar{C}_1 & \bar{A}_{22} \end{bmatrix} \\ \bar{K}_1 &= \begin{bmatrix} \bar{K}_{1,a} \\ \bar{K}_{1,b} \end{bmatrix}, \quad \bar{K}_2 = \begin{bmatrix} -\bar{K}_{1,a} \\ \bar{A}_{21}\bar{C}_1^{-1} - \bar{K}_{1,b} \end{bmatrix}, \quad \bar{K} = \begin{bmatrix} 0 \\ \bar{A}_{21}\bar{C}_1^{-1} \end{bmatrix} \end{aligned}$$

where $\bar{F}$ is Hurwitz stable by design of $\bar{K}_1$. Indeed, for $\bar{H} \doteq [\bar{H}_1^\top, \bar{H}_2^\top]^\top$ to satisfy (A.7):

$$\begin{bmatrix} \bar{H}_1\bar{C}_1 - I_n & 0 \\ \bar{H}_2\bar{C}_1 & -I_{n-q} \end{bmatrix} \begin{bmatrix} \bar{E}_1 \\ 0 \end{bmatrix} = 0$$

must hold, which, given invertibility of $\bar{E}_1$, implies:

$$\begin{cases} \bar{H}_1\bar{C}_1 = I_q & \implies \bar{H}_1 = \bar{C}_1^{-1}; \\ \bar{H}_2\bar{C}_1 = 0 & \implies \bar{H}_2 = 0. \end{cases} \quad (\text{A.27})$$

The definition of all other matrices follows.

Note that the pair $(\bar{F}, \bar{C})$ is not observable, and that all eigenvalues of $\bar{A}_{22}$ are the unobservable modes of the system. These are assumed to be stable, given condition ii. in Theorem A.5. The state $\bar{z}$ can therefore be written as $\bar{z} = [\bar{z}_1^\top, \bar{z}_2^\top]^\top$, where $\bar{z}_1$ and $\bar{z}_2$ are respectively the observable and unobservable portions of the state.

Recalling the structure of the UIO in (A.6):

$$\begin{cases} \dot{\bar{z}} = \bar{F}\bar{z} + \bar{K}\bar{y} \\ \hat{\bar{x}} = \bar{z} + \bar{H}\bar{y} \\ \hat{\bar{y}} = \bar{C}\bar{z} + \bar{y} \end{cases} \quad (\text{A.28})$$

it is evident, given the structure of the UIO matrices above, that $\bar{y}$ only influences the unobservable part of the state of the observer, $\bar{z}_2$. Furthermore, given stability of $-\bar{K}_{1,a}\bar{C}_1$,

$$\lim_{t \rightarrow \infty} \bar{z}_1(t) = \lim_{t \rightarrow \infty} e^{-\bar{K}_{1,a}\bar{C}_1 t} \bar{z}_1(0) = 0, \quad (\text{A.29})$$

and therefore $\hat{\bar{y}} \rightarrow \bar{y}$ asymptotically, irrespective of $f \neq 0$. As such $\lim_{t \rightarrow \infty} \bar{r} = 0$, the residual $r = Q(\bar{y} - \hat{\bar{y}}) = Q\bar{C}_1\bar{z}_1 \rightarrow 0$, independently of the presence of a process fault $f_n \neq 0$ or sensor fault $f_p \neq 0$. ■

As a consequence of Lemma A.1, if the number of unknown inputs from which the state estimate must be decoupled from is the same as the number of measurements, then it is not possible to detect an attack. This can intuitively be understood as the UIO not having sufficient redundancy to exploit to then be able to detect the fault. Indeed, for $p = q$, the UIO is such that it asymptotically tracks the output, and there is no further component that can be exploited that is not affected directly by an unknown input.

A.2.4 Augmented UIO for detection

Having shown that if $\text{rank } C = \text{rank } E$ a UIO cannot detect faults (and therefore attacks – as is the case in Chapter 3), we now present a method to enhance the UIO, such that it is made capable of detecting faults. We note that this solution is not possible for all scenarios, but is useful for that considered in Chapter 3. Specifically, for the purpose of this section, consider that the controller $\mathcal{C}$ for system (A.20) and diagnoser $\mathcal{D}$ exploiting the UIO (A.6) are not colocated, and that y and u are communicated without alteration (malicious or otherwise) from $\mathcal{C}$ to $\mathcal{D}$ for detection purposes.

Furthermore, assume that $\mathcal{C}$ has access to some knowledge of a portion of the unknown-input vector d , i.e. that d can be rearranged via a permutation into $\mathbf{d} = [d_1^\top, d_2^\top]^\top$, where d_1 is known to $\mathcal{C}$, but unknown to $\mathcal{D}$, and d_2 is unknown to both¹ $\mathcal{C}$ and $\mathcal{D}$.

We now introduce the following augmented state variable $\mathbf{x} \doteq [x_{art}^\top, e_{art}^\top]^\top$, where $x_{art} \in \mathbb{R}^n$ is some artificial state constructed by $\mathcal{C}$, the dynamics of which are simulated, and $e_{art} \doteq x - x_{art}$. Specifically, suppose the artificial state has linear dynamics:

$$\dot{x}_{art} = A_{art}x_{art} + E_1d_1, \quad (\text{A.30})$$

where E_1 is to be defined. By construction, we can reconstruct the state and output of the system (A.3) via the relation:

$$x = \begin{bmatrix} I_n & I_n \end{bmatrix} \mathbf{x} \quad . \quad (\text{A.31})$$

The dynamics of the augmented state and output can then be written as:

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{E}\mathbf{d} \\ \mathbf{y} = \mathbf{C}\mathbf{x} \end{cases} \quad (\text{A.32})$$

where the matrices are defined by:

$$\mathbf{A} \doteq \begin{bmatrix} A_{art} & 0 \\ A - A_{art} & A \end{bmatrix}, \quad \mathbf{E} \doteq \begin{bmatrix} E_1 & 0 \\ 0 & E_2 \end{bmatrix}, \quad \mathbf{C} \doteq \begin{bmatrix} I & 0 \\ C & C \end{bmatrix}. \quad (\text{A.33})$$

Note that x_{art} is completely included in augmented output $\mathbf{y}$, as it is computed directly by $\mathcal{C}$, and is therefore fully known. Nonzero matrices E_1, E_2 are constructed such that

$$\begin{bmatrix} E_1 & E_2 \end{bmatrix} = E \quad (\text{A.34})$$

¹This is indeed the case in Chapter 3, as the controller and the diagnoser are in two different subsystems, i.e. $\mathcal{C}_j$ and $\mathcal{D}_i$, $j \in \mathcal{N}_i$. In this case, $\mathcal{C}_j$ has access to inputs of the subsystem's dynamics unavailable to $\mathcal{D}_i$.

up to some column permutations of E . Thus,

$$\begin{bmatrix} E_1 & E_2 \end{bmatrix} \mathbf{d} = E\mathbf{d}. \quad (\text{A.35})$$

Additionally, E_1 and E_2 are selected such that the following hold:

$$\text{rank } E_1 < n, \quad \text{rank } E_2 < q \quad (\text{A.36})$$

and satisfy the inclusion

$$\text{Im } E_1 \subseteq \text{Im } E, \quad \text{Im } E_2 \subseteq \text{Im } E. \quad (\text{A.37})$$

Lemma A.2 ([GTB⁺20, Lem. 2]). Consider system (A.3); if (A, C, E) are such that conditions in Theorem A.5 hold, then $(\mathbf{A}, \mathbf{C}, \mathbf{E})$ also satisfy them.

If, additionally, $\text{rank } C = \text{rank } E$, then $\text{rank } \mathbf{C} > \text{rank } \mathbf{E}$. $\square$

Proof. From definition of $\mathbf{E}$ and $\mathbf{C}$, if $\text{rank}(CE) = \text{rank } E$, then $\text{rank}(\mathbf{CE}) = \text{rank}(\mathbf{E})$. Indeed:

$$\mathbf{CE} = \begin{bmatrix} E_1 & 0 \\ CE_1 & CE_2 \end{bmatrix} \quad (\text{A.38})$$

which, being block triangular, is such that:

$$\text{rank}(\mathbf{CE}) \geq \text{rank}(E_1) + \text{rank}(CE_2) = \text{rank}(E_1) + \text{rank}(E_2) = \text{rank}(E). \quad (\text{A.39})$$

Hence, given that $\text{rank}(\mathbf{CE}) \leq \min\{\text{rank } C, \text{rank } E\}$, it follows that $\text{rank}(\mathbf{CE}) = \text{rank } \mathbf{E}$.

To show that, if condition ii. in Theorem A.5 holds for (A, E, C) then it holds for the augmented system, note that a block-diagonal $\mathbf{S} \doteq \text{diag}(S_1, S_2)$ can be found such that

$$\mathbf{SE} = 0$$

as there exist solutions to both $S_1 E_1 = 0$ and $S_2 E_2 = 0$, following from properties (A.36).

Hence

$$\mathbf{SA} = \begin{bmatrix} S_1 A_{art} & 0 \\ * & S_2 A \end{bmatrix}$$

where $*$ represents the remaining term. Pair $(\mathbf{C}, \mathbf{SA})$ is detectable, as:

$$\text{rank} \begin{bmatrix} sI - \mathbf{SA} \\ \mathbf{C} \end{bmatrix} = \text{rank} \begin{bmatrix} sI - S_1 A_{art} & 0 \\ * & sI - S_2 A \\ I & 0 \\ C & C \end{bmatrix} = n + \text{rank} \begin{bmatrix} sI - S_2 A \\ C \end{bmatrix} \quad (\text{A.40})$$

which, for $S_2 \doteq S$ and H defined as in (A.15), is such that

$$\text{rank} \begin{bmatrix} sI - \mathbf{SA} \\ \mathbf{C} \end{bmatrix} = 2n, \quad \forall s \in \mathbb{C}_0^+,$$

given (C, SA) is a detectable pair by hypothesis.

The second part of the proposition holds, given that:

$$\text{rank } \mathbf{C} = n + \text{rank } C > \text{rank } E_1 + \text{rank } E_2 \quad (\text{A.41})$$

given $\text{rank } E_1 < n$ by construction and $\text{rank } C = \text{rank } E_2$ by hypothesis. ■

A.3 Minimum-variance unbiased estimator in the presence of unknown inputs and stochastic uncertainty

Having presented unknown-input observers and their properties, in this section, we summarise the results presented in [GDM07]. The objective of the paper is to design a minimum-variance unbiased (MVU) estimate capable of estimating the state and unknown inputs of a linear system. The summary reported here is added to give additional explanation for the decentralised estimation framework in Chapter 4 in the context of distributed cyber-attack detection of large-scale interconnected systems with stochastic uncertainties. Although we provide some remarks

summarising the insight offered in [GDM07], we refer the interested reader to the paper for additional discussion of results, and for a comparison with other unbiased estimators in the literature, such as [Kit87, DZ97, Hsi00].

We highlight that, while in [GDM07] the authors deal with a time-varying discrete time system, here, for clarity of presentation, we consider an LTI system. Specifically, we consider the following dynamics with unknown inputs and subject to stochastic uncertainties:

$$\begin{cases} x(k+1) = Ax(k) + Gd(k) + w(k) \\ y(k) = Cx(k) + v(k) \end{cases} \quad (\text{A.42})$$

where $x \in \mathbb{R}^n$ is the state of the system, $d(k) \in \mathbb{R}^q$ is an unknown input, and $y(k) \in \mathbb{R}^p$ is the output. We disregard any possible effect of an input signal on the system, as so long as it is known it can be accounted for and does not affect the properties of the estimator. The vectors $w(k)$ and $v(k)$ model process and measurement uncertainties, and are modelled as independent identically distributed Gaussian processes

$$w(k) \sim \mathcal{N}(0, W), \quad v(k) \sim \mathcal{N}(0, V), \quad \forall k \in \mathbb{N}_0, \quad (\text{A.43})$$

with $W \succcurlyeq 0$ and $V \succcurlyeq 0$. It is assumed that

$$\text{Cov}[w(k), v(\kappa)] = \mathbf{E}[w(k)v(\kappa)^\top] = 0, \quad \forall k, \kappa \in \mathbb{N}_0.$$

Furthermore, initial condition $x(0) = x_0$ is assumed to be a realization of the Gaussian distribution with mean $\bar{x}_0$ and variance Π_0 :

$$x_0 \sim \mathcal{N}(\bar{x}_0, \Pi_0), \quad (\text{A.44})$$

independent of $w(k), v(k)$, and assume that (A, C) is an observable pair.

Assumption A.1. For system in (A.42), $\text{rank}CG = \text{rank}G = q$. ◁

Remark A.1. Note that this is also one of the necessary and sufficient conditions for the

construction of a UIO in a deterministic system, as expressed in Theorem A.5. $\triangleleft$

Consider the recursive filter of the form:

$$\hat{x}(k|k-1) = A\hat{x}(k-1|k-1) \quad (\text{A.45a})$$

$$\hat{d}(k-1|k) = M(k) (y(k) - C\hat{x}(k|k-1)) \quad (\text{A.45b})$$

$$\hat{x}^*(k|k) = \hat{x}(k|k-1) + Gd(k-1|k) \quad (\text{A.45c})$$

$$\hat{x}(k|k) = \hat{x}^*(k|k) + K(k) (y(k) - C\hat{x}^*(k|k)) \quad (\text{A.45d})$$

where $M(k)$ and $K(k)$ are to be determined. Note that, each step in (A.45) represents a different stage in the estimation process: $\hat{x}(k|k-1)$ is a one-step ahead prediction starting from unbiased estimate $\hat{x}(k-1|k-1)$, biased by lack of knowledge of the unknown input; $\hat{d}(k-1|k)$ is the estimate of the unknown input, where $M(k)$ is defined such that $\hat{d}$ is an MVU estimate of the unknown input; $\hat{x}^*(k|k)$ is an updated prediction of the state of the system, which is unbiased; finally $\hat{x}(k|k)$ is the filtered estimate of the state $x(k)$, for which the variance is minimised defining $K(k)$ similarly to the Kalman filter.

Note that $\hat{d}$ is delayed, as it relies on $y(k) - Cx(k|k-1)$ to reconstruct information on the unknown input $d(k-1)$. This is to be expected, given d influences the state of the system dynamically.

Define the biased prediction innovation $\tilde{r}(k) \doteq y(k) - C\hat{x}(k|k-1)$, and estimation error $\epsilon(k) = x(k) - \hat{x}(k|k)$. From (A.42) and (A.45a)

$$\tilde{r}(k) = CGd(k-1) + \eta(k) \quad (\text{A.46})$$

with

$$\eta(k) = C(A\epsilon(k-1) + w(k-1)) + v(k). \quad (\text{A.47})$$

Supposing for the time being that $\hat{x}(k-1|k-1)$ is unbiased, then $\mathbf{E}[\epsilon(k-1)] = 0$, $\mathbf{E}[\eta(k)] = 0$ and thus the expected value of the biased prediction innovation is given by:

$$\mathbf{E}[\tilde{r}(k-1|k)] = CGd(k-1). \quad (\text{A.48})$$

As such, an unbiased estimate of $d(k-1)$ can be obtained from $\tilde{r}(k)$. In the following Theorem a necessary and sufficient condition is given for the unbiasedness of estimate $\hat{d}(k-1|k)$.

Theorem A.6 ([GDM07, Th.1]). Let $\hat{x}(k-1|k-1)$ be unbiased, then (A.45a)-(A.45b) is an unbiased estimate of $d(k-1)$ if and only if

$$M(k)CG = I_q, \quad \forall k \in \mathbb{N}_0. \quad (\text{A.49})$$

□

In order to obtain a minimum-variance estimate from $\tilde{r}$, the matrix $M(k)$ is obtained through application of a weighted least squares estimate:

$$M(k) = (F^\top H(k)^{-1} F)^{-1} F^\top H(k)^{-1} \quad (\text{A.50})$$

using weighting $H(k)^{-1} \doteq \text{Var}[\eta(k)]^{-1}$. Specifically, in (A.50), $F \doteq CG$, and $H(k)$ is calculated from (A.47) as

$$H(k) = C (A\Pi(k-1)A^\top + W) C^\top + V \quad (\text{A.51})$$

where $\Pi(k-1) \doteq \text{Var}[\epsilon(k-1)]$, to be defined explicitly in what follows.

Let us now present some of the main theorems in [GDM07], demonstrating that $\hat{x}(k|k)$ and $\hat{d}(k-1|k)$ are in fact minimum-variance unbiased estimates of $x(k)$ and $d(k-1)$, respectively.

Theorem A.7 ([GDM07, Th.2]). Let Assumption A.1 hold, let $\hat{x}(k-1|k-1)$ be unbiased, let $H(k) \succ 0$ and define $M(k)$ as in (A.50). Then (A.45b) is the MVU estimator of $d(k-1)$, given innovation $\tilde{r}(k)$. The variance of $\hat{d}(k-1|k)$ is given by $(F^\top H^{-1} F)^{-1}$. □

Remark A.2. Note that Theorem A.7 makes use of a weighted least squares estimate to minimise variance of $\hat{d}$ from (A.46), by normalising the variance of e in (A.46). Therefore, $\hat{d}$ is an MVU estimate of d . ◁

Theorem A.8 ([GDM07, Th.4]). Let $\hat{x}(k-1|k-1)$ and $\hat{d}(k-1|k)$ be unbiased estimates of $x(k-1)$ and $d(k-1)$, then (A.45c)–(A.45d) are unbiased estimates of $x(k)$, irrespective of $K(k)$. □

Theorem A.9 ([GDM07, Th.12]). Consider a joint input and state estimator of the recursive form (A.45). Let $M(k)$ be defined as in (A.50) and let $K(k)$ be given by

$$K(k) = (A\Pi(k-1)A^\top + W)CH^{-1}, \quad (\text{A.52})$$

then (A.45b) and (A.45d) are unbiased estimators of $d(k-1)$ and $x(k)$ minimising the mean square error over the class of all linear unbiased estimates based on $\hat{x}(0|0)$ and $\mathcal{Y}(k) \doteq \{y(0), y(1), \dots, y(k)\}$. $\square$

Remark A.3. The authors of [GDM07] point out that the definition of $K(k)$ in (A.52) makes the estimator (A.45) have the same state update as in [Kit87] [see [GDM07, Th.5]]. However, the proof of optimality in their work establishes a stronger connection to the Kalman filter, while also proving the optimality of the unknown input estimate (A.45b). $\triangleleft$

While we omit the derivation of the proof of optimality of either of the estimates in (A.45), we derive the variance $\Pi(k)$ of the estimation error $\epsilon(k)$. Indeed, to generate an unbiased minimum estimate of the state, once $\Pi(k)$ is formulated, the condition on $K(k)$ is defined such that $\hat{x}(k|k)$ in (A.45d) is MVU. Start by defining $\epsilon^*(k) \doteq x(k) - \hat{x}^*(k|k)$:

$$\epsilon^*(k) = A\epsilon(k-1) + G\delta(k-1) + w(k-1) \quad (\text{A.53})$$

where $\delta \doteq d - \hat{d}$ is the unknown-input estimation error. By direct computation:

$$\begin{aligned} \delta(k-1) &= (I_q - M(k)CG)d(k-1) - M(k)\eta(k) \\ &= -M(k)\eta(k) \end{aligned}, \quad (\text{A.54})$$

given definition of $M(k)$ satisfying (A.49). Note that, given definition of $M(k)$ in (A.50), $\Delta(k-1) \doteq \text{Var}[\delta(k-1)]$

$$\Delta(k-1) = (F^\top H^{-1}F)^{-1},$$

the same as $\hat{d}(k-1|k)$. Thus, recalling definition of η in (A.47), $\epsilon^*(k)$ can be rewritten as

$$\epsilon^*(k) = (I_n - GM(k)C)(A\epsilon(k-1) + w(k-1)) - GM(k)v(k), \quad (\text{A.55})$$

with variance $\Pi^*(k) \doteq \text{Var}[\epsilon^*(k)]$:

$$\Pi^*(k) = (I_n - GM(k)C)(A\Pi(k-1)A^\top + W)(I_n - GM(k)C)^\top + GM(k)VM(k)^\top G^\top. \quad (\text{A.56})$$

All the elements necessary to compute $\Pi(k)$ are now available. Given the definition of the estimate in (A.45d), the estimation error $\epsilon(k) = x(k) - \hat{x}(k|k)$ can be found to be:

$$\epsilon(k) = (I_n - K(k)C)\epsilon^*(k) - K(k)v(k) \quad (\text{A.57})$$

which, given formulation of ϵ^* in (A.55), can be rewritten as:

$$\epsilon(k) = (I_n - K(k)C)(I_n - GM(k)C)[A\epsilon(k-1) + w(k-1)] - [(I_n - K(k)C)GM(k) + K(k)]v(k). \quad (\text{A.58})$$

Remark A.4. The authors of [GDM07] note that the definition of the estimation error in (A.58), after some manipulations, has close connection to the Kalman filter. Indeed, it can be seen as the dynamical evolution of the state estimation error for the system (A^*, C) , where

$$A^* = (I_n - GM(k)C)A$$

and process noise $w^*(k) = (I_n - GM(k)C)w(k-1) - GM(k)v(k)$, correlated to measurement noise with covariance $\text{Cov}[w^*(k), v(k)] = -GM(k)V$. Hence, the remainder of the discussion around design of $K(k)$ can be seen as a standard Kalman filter design. $\triangleleft$

Let us now conclude by defining the estimation error covariance matrix $\Pi(k)$, while referring the interested reader to [GDM07] for further discussion on the definition of $K(k)$ such that $\hat{x}(k|k)$ is a minimum-variance estimate of $x(k)$.

It follows from definition of $\epsilon(k)$ and of $\Pi^*(k)$ that the error covariance matrix is given by:

$$\begin{aligned} \Pi(k) = & K(k)\tilde{R}^*(k)K(k)^\top + \Pi^*(k) - \Pi^*(k)C^\top K(k)^\top - K(k)C^\top \Pi^*(k) \\ & + GM(k)VK(k)^\top + K(k)VM(k)^\top G^\top \end{aligned} \quad (\text{A.59})$$

where $\tilde{R}^*(k) \doteq \text{Var}[\tilde{r}^*(k)]$ is the variance of the updated innovation $\tilde{r}^* \doteq y(k) - C\hat{x}^*(k|k)$, derived from (A.55) as:

$$\begin{aligned}\tilde{R}^*(k) &= C\Pi^*(k)C^\top + V - CGM(k)V - VM(k)^\top G^\top C^\top \\ &= (I_p - CGM(k))H(I_p - CGM(k))^\top.\end{aligned}\tag{A.60}$$

thus concluding our discussion of the statistical properties of the decentralised filter (A.45).

A.4 Conclusion

In this appendix we have summarised some of the main results on unknown input observability and reconstruction, unknown-input observers, and unbiased minimum variance filters, which have been exploited throughout the thesis.

Appendix B

Discretisation of distributed generation unit dynamics

Let us in this appendix present some details on how we discretise the dynamics of a DC microgrid composed of networked DGUs are defined. For clarity of presentation, we repropose the circuit diagram of a DGU in Figure B.1.

First, let us recall the discussion on the state space dynamics in Chapter 2. The dynamics of each DGU are given by the following linear time-invariant system:

$$\begin{cases} \dot{x}_i = A_{ii}x_i + B_i u_i^{\text{sw}} + \sum_{j \in \mathcal{N}_i} A_{ij}x_j + E_i d_i \\ y_i = C_i x_i \end{cases} \quad (\text{B.1})$$

where $x_i = [V_i, I_{ti}, \nu_i]^\top$ is the state of the system, with u_i^{sw} the switching control input, $d_i = \begin{bmatrix} I_{Li} & V_{ref,i} \end{bmatrix}^\top$ the exogenous input composed of the local load current and voltage reference. The matrices are defined in (2.32) in Chapter 2. The switching control input u_i^{sw} is defined as in (2.30):

$$u_i^{\text{sw}}(t) = \begin{cases} V_{si}, & t \in [kT_i^{\text{sw}}, kT_i^{\text{sw}} + \delta_i(kT_i^{\text{sw}})T_i^{\text{sw}}) \\ 0, & t \in [kT_i^{\text{sw}} + \delta_i(kT_i^{\text{sw}})T_i^{\text{sw}}, (k+1)T_i^{\text{sw}}) \end{cases} \quad (\text{B.2})$$

where V_{si} is the source voltage of the i -th DGU, T^{sw} is the considered switching period of the

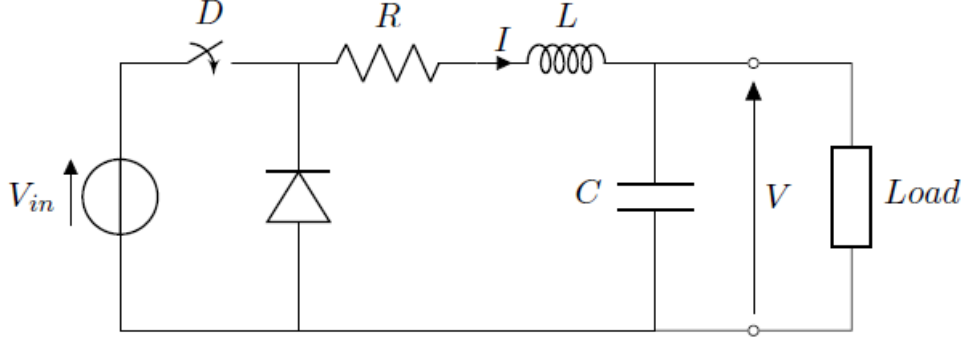


Figure B.1: Circuit diagram of Buck converter in series with an RLC filter and a resistive load. The switch D represents a transistor, such as an IGBT.

Buck DC-DC converter, and the duty cycle $\delta_i(\cdot)$ is computed from the primary control loop:

$$\delta_i(t) = \frac{1}{V_{si}} K_i x_i(kT_i^{\text{sw}}), \quad t \in [kT_i^{\text{sw}}, (k+1)T_i^{\text{sw}}) \quad (\text{B.3})$$

with $K_i \doteq [k_{i,1}, k_{i,2}, k_{i,3}]$. The local voltage reference $V_{ref,i}$ is dependent on the secondary control loop, and therefore defined as:

$$V_{ref,i} = V_{ref} + \Delta V_i \quad (\text{B.4})$$

where V_{ref} is the voltage reference common to all DGUs [Assumption 2.3], while ΔV_i is the result of the consensus protocol (2.12).

Recalling our discussion in the introduction to Chapter 4, we discretise the dynamics of each DGU in the DC microgrid independently of its neighbours using exact discretisation, and given a common, synchronous, sampling period T_s . This is equivalent to the following assumption.

Assumption B.1. For all $t \in [kT_s, (k+1)T_s)$, all inputs u_i^{sw}, d_i and $x_j, j \in \mathcal{N}_i$ are constant, for all DGU $i, i \in \mathcal{N}$. $\triangleleft$

Furthermore, we consider that the sampling and switching instances coincide, and therefore that sampling period $T_s = T^{\text{sw}}$. The discrete time dynamics are approximated from (B.1) as:

$$\begin{aligned} x_i(k+1) &= A_{ii}^d x_i(k) + B_i^d(k) \bar{u}_i + \sum_{j \in \mathcal{N}_i} A_{ij}^d x_j(k) + E_i^d d_i \\ y_i(k) &= C_i x_i(k) \end{aligned} \quad (\text{B.5})$$

with constant $\bar{u}_i = V_{si}$. The matrices $A_{ii}^d, E_i^d, A_{ij}^d$ in (B.5) are defined as the following time

invariant matrices:

$$A_{ii}^d \doteq e^{A_{ii}T_s}, \quad E_i^d \doteq \left(\int_0^{T_s} e^{A_{ii}\tau} d\tau \right) E_i, \quad A_{ij}^d \doteq \left(\int_0^{T_s} e^{A_{ii}\tau} d\tau \right) A_{ij} \quad (\text{B.6})$$

while the input matrix is defined as the result of the following time-varying expression

$$B_i^d(k) = e^{A_{ii}\delta^*(k)T_s} \int_0^{\delta(k)T_s} e^{A_{ii}\tau} d\tau B_i. \quad (\text{B.7})$$

This is done to avoid including the switching error in the discrete-time dynamics, with $\delta_i(k)$ defined as in (2.22). The time-varying definition of $B_i^d(k)$ can be derived by recalling the solution of the state under switching dynamics given in (2.26). Indeed, recall that at time $t = T_s$, with initial conditions $x_i(0)$, the state of the DGU without exogenous inputs $x_j, j \in \mathcal{N}_i$ and d_i following dynamics (2.14) is

$$x_i(T_s) = e^{A_{ii}T_s} x_i(0) + \int_0^{\delta(k)T_s} e^{A_{ii}(T_s-\tau)} B_i V_{si} d\tau. \quad (\text{B.8})$$

The definition of $B_i^d(k)$ follows from its definition in (B.7).

Appendix C

Augmenting DGU state space for UIO detectability

In this appendix, we analyse the possibility of constructing an unknown-input observer (UIO) for the case-study in Chapter 3, where a network of N DGUs in a DC microgrid is considered. Indeed, following from the design of the diagnoser $\mathcal{D}_i$ for each DGU i , a bank of $N_i = |\mathcal{N}_i|$ UIOs are to be designed for each of the neighbours in the set $\mathcal{N}_i$. Given the definition of the subsystem dynamics with unknown inputs in (3.54), the unknown input vector $\check{d}_i$ defined in (3.56) in the case of DGUs, in light of their dynamics defined in Chapter 2, is:

$$\check{d}_i = \left[V_{ti}, I_{Li}, V_{ref,i}, V_{j_1}, \dots, V_{j_{N_i}} \right]^\top \quad (\text{C.1})$$

and the matrix $\bar{E}_i \in \mathbb{R}^{n_i \times q_i}$ is defined such that

$$\text{Im}[\bar{E}_i] = \text{Im} \left[B_i, E_i, A_{ij_1}, \dots, A_{ij_{N_i}} \right]. \quad (\text{C.2})$$

with $\{j_1, \dots, j_{N_i}\} \equiv \mathcal{N}_i$. From the definition of $\bar{E}_i$ and that of matrices in (2.32) in Chapter 2, it is clear that

$$\text{rank}(\bar{E}_i) = q_i = p_i = \text{rank}(C_i), \quad (\text{C.3})$$

thus making any UIO constructed with dynamics (2.31) unable to detect attacks on y_i , as shown in Lemma A.1, in Appendix A.2.3. It is therefore necessary to augment the system with some

artificial dynamics such that UIOs may detect attacks. Here, we choose the average dynamics of the DGU introduced in Chapter 2.1.3 to augment the state.

In the following we construct the augmented system and demonstrate how this choice is appropriate, i.e. that for the augmented system $\bar{p}_i > \bar{q}_i$, where $\bar{p}_i$ is the rank of the augmented system output, while $\bar{q}_i$ is the augmented system unknown input.

Recalling our discussion on the average dynamics of the DGU in Chapter 2.1.3, the average input to the system is defined as:

$$u_i^{\text{avg}}(t) = \delta_i(t)V_{si} = K_i x_i \quad (\text{C.4})$$

and therefore the dynamics of the average state x_i^{avg} and error $\epsilon_i^{\text{avg}} \doteq x_i - x_i^{\text{avg}}$:

$$\dot{x}_i^{\text{avg}} = A_{ii}x_i^{\text{avg}} + B_i u_i^{\text{avg}} + E_i d_i \quad (\text{C.5})$$

$$\dot{\epsilon}_i^{\text{avg}} = -B_i K_i x_i^{\text{avg}} + (A_{ii} - B_i K_i)\epsilon_i^{\text{avg}} + B_i u_i^{\text{sw}} + \sum_{j \in \mathcal{N}_i} A_{ij} x_j \quad (\text{C.6})$$

grouping all of the “known-unknown inputs” (i.e. those inputs that are known to the local controller but unknown to its neighbours) d_i in the average state dynamics, while the switching input u_i^{sw} and the “unknown-unknown inputs” $x_{j \in \mathcal{N}_i}$ in the average error. Note that, by definition, the original state can be reconstructed, as $x_i = x_i^{\text{avg}} + \epsilon_i^{\text{avg}}$. Hence, the augmented state $\mathbf{x}_i \doteq \begin{bmatrix} x_i^{\text{avg}\top}, \epsilon_i^{\text{avg}\top} \end{bmatrix}^\top$ has dynamics:

$$\begin{cases} \dot{\mathbf{x}}_i = \mathbf{A}_{ii}\mathbf{x}_i + \mathbf{E}_i \mathbf{d}_i \\ \mathbf{y}_i = \mathbf{C}_i \mathbf{x}_i \end{cases} \quad (\text{C.7})$$

with $\mathbf{d}_i \doteq [d_i^\top, u_i^{\text{sw}\top}, x_{j \in \mathcal{N}_i}^\top]^\top$, $\mathbf{y}_i \doteq [x_i^{\text{avg}\top}, y_i^\top]^\top$, and matrices

$$\mathbf{A}_{ii} \doteq \begin{bmatrix} A_{ii} + B_i K_i & B_i K_i \\ -B_i K_i & A_{ii} - B_i K_i \end{bmatrix}, \quad \mathbf{E}_i \doteq \begin{bmatrix} E_i & 0 & 0 \\ 0 & B_i & A_{ij, j \in \mathcal{N}_i} \end{bmatrix}, \quad \mathbf{C}_i \doteq \begin{bmatrix} I_{n_i} & 0 \\ C_i & C_i \end{bmatrix}. \quad (\text{C.8})$$

In order to demonstrate that (C.7) is an appropriate augmented system to construct a UIO

which can detect attacks, we prove that the following holds.

Lemma C.1. Consider a DGU with dynamics defined in Chapter 2. Suppose that the physical dynamics are augmented with some auxiliary dynamics, as in (C.7). Then, the following hold:

1. $\text{rank}(\mathbf{C}_i \mathbf{E}_i) = \text{rank}(\mathbf{E}_i)$;
2. $\bar{p}_i > \bar{q}_i$, where $\bar{p}_i = \text{rank } \mathbf{C}_i$, $\bar{q}_i = \text{rank } \mathbf{E}_i$.

Therefore, it is possible to construct a UIO as in Appendix A, such that it is capable of detecting attacks. $\square$

Proof. The first condition follows directly from definition of $\mathbf{C}_i$ and C_i , given $C_i = I_3$ for all DGUs from the discussion in Chapter 2.

To prove the second condition, recalling that $n_i = 3$ by definition of the DGU dynamics in Section 2.1.4, $\bar{p}_i = 6$ by definition of $\mathbf{C}_i$, while from (2.32) $\bar{q}_i = 4$. Furthermore, given their definitions in Chapter 2, each block diagonal element of $\mathbf{C}_i$, i.e. $\mathbf{C}_{i,k}$, $k \in \{1, 2\}$, is such that

$$\text{rank}(\mathbf{C}_{i,k}) > \text{rank}(\mathbf{E}_{i,k}) \quad k \in \{1, 2\} \quad (\text{C.9})$$

with $\mathbf{E}_{i,k}$ the k -th block diagonal element of $\mathbf{E}_i$. Thus, a UIO capable of detecting attacks may be constructed from the augmented dynamics (C.7). $\blacksquare$

We have therefore designed appropriate augmented dynamics for the i -th DGU, enabling a UIO to be constructed which does not satisfy condition in Lemma A.1.