# Control-Enabled Approaches for Active Detection of Cyberattacks on Process Control Systems

By

Shilpa Narasimhan

Dissertation

Submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy

in

Chemical Engineering

in the

Office of Graduate Studies

of the

University of California

Davis

Approved:

Matthew J. Ellis, Chair

Nael H. El-Farra, Co-chair

Ahmet N. Palazoglu

Committee in Charge

2024

*To all my teachers.*

# Contents

List of Figures . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . vii

List of Tables . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . xiv

Abstract . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . xv

Acknowledgments . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . xviii

**1 Introduction** **1**

1.1 Background and Motivation . . . . . . . . . . . . . . . . . . . . . . . . 1

1.1.1 Cyberattacks on Critical Infrastructure . . . . . . . . . . . . . . . 1

1.1.2 False Data Injection Cyberattacks . . . . . . . . . . . . . . . . . . 2

1.1.3 Control-Enabled Approaches for Cyberattack Resilience . . . . . . 4

1.1.4 Detection of Cyberattacks on PCSs . . . . . . . . . . . . . . . . . . 6

1.2 Objectives and Organization of the Dissertation . . . . . . . . . . . . . 7

**2 Detectability-Based Controller Design Screening for Processes Under Multiplicative Cyberattacks** **12**

2.1 Preliminaries . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

2.1.1 Notation and Definitions . . . . . . . . . . . . . . . . . . . . . . . 13

2.1.2 Process Model and Control System . . . . . . . . . . . . . . . . . 13

2.1.3 Residual-Based Attack Detection Scheme . . . . . . . . . . . . . . 15

2.2 Controller Screening Methodology . . . . . . . . . . . . . . . . . . . . . 18

2.2.1 Detectability-Based Classification of Attacks . . . . . . . . . . . . 18

2.2.2 Minimum Invariant Set and Residual Bound Estimation . . . . . 22

2.2.3 Controller Screening Methodology . . . . . . . . . . . . . . . . . . 27

2.3 Application to Illustrative Processes . . . . . . . . . . . . . . . . . . . 29

2.3.1 Illustrative Process 1: Scalar Process . . . . . . . . . . . . . . . . 29

2.3.2 Illustrative Process 2: Continuous Stirred Tank Reactor . . . . . . 33

2.4 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 39

**3 Active Multiplicative Cyberattack Detection Utilizing Controller Switching for Process Systems 40**

3.1 Preliminaries . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 41

3.1.1 Notation and Definition . . . . . . . . . . . . . . . . . . . . . . 41

3.1.2 Class of Processes and Control System Design . . . . . . . . . . . 41

3.2 Active Multiplicative Attack Detection Utilizing Controller Switching . . 44

3.2.1 Residual-Based Detection Scheme and Attack Detectability . . . . 44

3.2.2 Active Attack Detection Methodology . . . . . . . . . . . . . . . 48

3.3 Application to a Chemical Process . . . . . . . . . . . . . . . . . . . . 57

3.3.1 Application of the Active Detection Methodology to the Linearized CSTR . . . . . . . . . . . . . . . . . . . . . . . . . . . . 59

3.3.2 Application of the Active Detection Methodology to the Nonlinear CSTR . . . . . . . . . . . . . . . . . . . . . . . . . . . . 63

3.4 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 71

**4 A Control-Switching Approach for Cyberattack Detection in Process Systems with Minimal False Alarms 72**

4.1 Preliminaries . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 73

4.1.1 Notation and Definitions . . . . . . . . . . . . . . . . . . . . . . 73

4.1.2 Class of Processes . . . . . . . . . . . . . . . . . . . . . . . . . 73

4.2 Class of Attack Detection Schemes . . . . . . . . . . . . . . . . . . . . 75

4.3 Active Detection Method . . . . . . . . . . . . . . . . . . . . . . . . 78

4.3.1 Controller Switching for Active Detection . . . . . . . . . . . . . 78

4.3.2 Confidence Region-Based Switching Condition to Eliminate False Alarms . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 80

4.3.3 Minimizing False Alarms . . . . . . . . . . . . . . . . . . . . . . 83

4.4 Illustrative Case Studies . . . . . . . . . . . . . . . . . . . . . . . . . 88

4.4.1 Application to a Scalar Process . . . . . . . . . . . . . . . . . . 88

4.4.2 Application to a Chemical Process . . . . . . . . . . . . . . . . . 95

4.5 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 103

**5 A Reachable Set-Based Scheme for the Detection of False Data Injection Cyberattacks on Dynamic Processes 105**

5.1 Preliminaries . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 106

5.1.1 Notation and Definitions . . . . . . . . . . . . . . . . . . . . . . 106

5.1.2 Class of Attack-Free Processes . . . . . . . . . . . . . . . . . . . 106

5.1.3 False Data Injection Attacks . . . . . . . . . . . . . . . . . . . . 108

5.2 Attack Detection for Processes During Transient Operation . . . . . . . . 109

5.2.1 Reachable Set-Based Detection Scheme . . . . . . . . . . . . . . 109

5.2.2 Classification of Attack Detectability . . . . . . . . . . . . . . . 112

5.3 Numerical Results: Scalar Process Example . . . . . . . . . . . . . . . 117

5.3.1 Application of the Reachable Set-Based Detection Scheme and Detectability-Based Classification of Attacks . . . . . . . . . . . . . . . . . . 118

5.3.2 Factors that Influence the Detectability of a Multiplicative False Data Injection Attack . . . . . . . . . . . . . . . . . . . . . . . . 120

5.3.3 Factors Influencing the Detectability of an Additive FDIA . . . . 123

5.4 Numerical Results: Chemical Process Example . . . . . . . . . . . . . 126

5.4.1 Application of the Reachable Set-Based Detection Scheme . . . . 128

5.4.2 Application of Detectability-Based Classification of an Attack . . 130

5.5 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 133

**6 Detection of Multiplicative False Data Injection Cyberattacks on Process Control Systems via Randomized Control Mode Switching 134**

6.1 Preliminaries . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 135

6.1.1 Notation and Definitions . . . . . . . . . . . . . . . . . . . . . . 135

6.1.2 Class of Processes . . . . . . . . . . . . . . . . . . . . . . . . . . 135

6.1.3 Class of Multiplicative False Data Injection Attacks . . . . . . . . 137

6.1.4 Class of Reachable Set-Based Detection Schemes . . . . . . . . . . 138

6.2 Randomized Control Mode Switching for Detection of Cyberattacks . . . 142

6.2.1 Control Mode Switching for Cyberattack Detection . . . . . . . . 142

6.2.2 Algorithms for the Randomized Control Mode Switching-Enabled Cyberattack Detection Method . . . . . . . . . . . . . . . . . . . . 148

6.3 Application of Randomized Control Mode Switching Algorithms to Illustrative Processes . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 154

6.3.1 Illustrative Scalar Process . . . . . . . . . . . . . . . . . . . . . . . . 154

6.3.2 A Continuous Stirred-Tank Reactor . . . . . . . . . . . . . . . . . . 161

6.4 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 173

**7 A Set-Based Control Mode Selection Approach for Active Detection of False Data Injection Cyberattacks 174**

7.1 Preliminaries . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 175

7.1.1 Notation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 175

7.1.2 Class of Processes . . . . . . . . . . . . . . . . . . . . . . . . . . . 175

7.1.3 Class of False Data Injection Attacks . . . . . . . . . . . . . . . . 177

7.1.4 Monitoring Variable and Set-Based Detection Scheme . . . . . . . 178

7.2 Active Detection for Stealthy Attacks . . . . . . . . . . . . . . . . . . . 179

7.2.1 Active Detection with Set Point and Control Parameter Changes 180

7.2.2 Selecting an Attack-Revealing Control Mode for Active Detection 182

7.3 Application to an Illustrative Process . . . . . . . . . . . . . . . . . . . 185

7.4 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 190

**8 Conclusions 191**

**A Proofs for Chapter 2 193**

**B Proofs for Chapter 3 194**

**C Proofs for Chapter 4 198**

**D Proofs for Chapter 5 202**

**E Proofs for Chapter 6 204**

# List of Figures

1.1 Block diagram of a PCS illustrating the potential targets for a false data injection cyberattack. . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

2.1 Example of an undetectable attack with $R_a \leq R$ but $D_{r_a} \not\subset D_r$. . . . . . . 21

2.2 The outer estimate ($D_{\xi}^{est}$) of the minimum invariant set ($D_{\xi}$) of the augmented closed-loop system. . . . . . . . . . . . . . . . . . . . . . . . 22

2.3 (a) An illustrative case showing false positive test result for undetectability with $R_a^{est} < R^{est}$, $R_a > R$. (b) An illustrative case showing false negative test result for undetectability with $R_a^{est} > R^{est}$, $R_a < R$. . . . . . . . . . . 25

2.4 Flowchart for the controller design screening algorithm. . . . . . . . . . . 28

2.5 The estimates of the minimum invariant set of (a) the attack-free augmented closed-loop system (${D_{\xi}}^{est}$) and (b) the attacked augmented closed-loop system ($D_{\xi_a}^{est}$) for the scalar process in the undetectable attack case. 31

2.6 The estimates of the minimum invariant set of (a) the attack-free augmented closed-loop system (${D_{\xi}}^{est}$) and (b) the attacked augmented closed-loop system ($D_{\xi_a}^{est}$) for the scalar process in the potentially detectable attack case. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 31

2.7 (a) The CUSUM statistic for 1000 simulations of the attacked scalar process for the potentially detectable attack case. (b) The number of threshold breaches per simulation for 1000 simulations of the attacked scalar process for the potentially detectable attack case. . . . . . . . . . . . . . . . . 32

2.8 The outer estimate of the residual set ($D_{r_a}^{est}$) of the attacked CSTR and the inner estimate of the residual set ($D_{r_e}^{est}$) of the attack-free CSTR in the undetectable attack case. . . . . . . . . . . . . . . . . . . . . . . . . 36

2.9 CUSUM statistic values observed over 1000 simulations of the attacked CSTR process for the undetectable attack case. . . . . . . . . . . . . . 37

2.10 The outer estimate of the residual set ($D_r^{est}$) of the attack-free CSTR and the inner estimate of the residual set ($D_{r_{a,e}}^{est}$) of the attacked CSTR in the potentially detectable attack case. . . . . . . . . . . . . . . . . . . . . . . 38

2.11 CUSUM statistic values observed over 1000 simulations of the attacked CSTR process for the potentially detectable attack case. . . . . . . . . . 38

3.1 (a) An example residual trajectory for the attack-free closed-loop process with the control system switch from nominal parameters to attack-sensitive parameters occurring at $t_s$. (b) An example residual trajectory for the attacked closed-loop process with the control parameter switch occurring at $t_s$ where the attack is detected at $t_d$. . . . . . . . . . . . . . . . . . . . . 49

3.2 Flowchart for the active attack detection methodology. . . . . . . . . . . 51

3.3 (a) An example showing the evolution of the augmented state trajectory for the attack-free closed-loop process with a control parameter switch generating zero false alarms. (b) An example showing the evolution of the augmented state trajectory for the attack-free closed-loop process with a control system parameter switch that may generate false alarms. . . . . . 53

3.4 An example showing the evolution of the augmented state trajectory for the attacked closed-loop process. After a control system parameter switch to attack-sensitive parameters, the augmented state mimics the transient behavior of an attack-free process briefly. . . . . . . . . . . . . . . . . . 54

3.5 The residual values of the attacked linearized CSTR process with (a) with nominal parameters before and after a switch to the attack-sensitive parameters and (b) with attack-sensitive parameters where the attack is detected at time $t_d = 2.74\,\mathrm{h}$. . . . . . . . . . . . . . . . . . . . . . . . . . 61

3.6 (a) The residual values of the attack-free nonlinear CSTR process with nominal parameters before and after a switch to the attack-sensitive parameters. (b) The residual values of the attack-free nonlinear CSTR process with attack-sensitive parameters. . . . . . . . . . . . . . . . . . . . . . 65

3.7 (a) The residual values of the attacked nonlinear CSTR process with nominal parameters before and after a switch to the attack-sensitive parameters. (b) The detection of the attack at the time $t_d = 2.69\,\text{h}$ after a switch from nominal parameters to attack-sensitive parameters. . . . . . . . . . . . . 66

3.8 (a) The CUSUM statistic for the attacked CSTR without the active detection methodology implemented. (b) The CUSUM statistic for the attacked CSTR process with the active detection methodology implemented, showing that the attack is detected at time $t = 2.69$ h. . . . . . . . . . . . . 69

3.9 (a) The residual values of the attacked nonlinear CSTR with nominal parameters before and after a switch to attack-sensitive parameters. (b) The residual values of the attacked nonlinear CSTR with attack-sensitive parameters showing attack detection at time $t_d = 2.69$ h. . . . . . . . . . 70

4.1 (a) Confidence regions for the attack-free process under the nominal mode at the time $t_s = 250$. (b) Confidence regions for the attack-free process under the attack-sensitive mode at the time $t_s = 350$. . . . . . . . . . . 91

4.2 (a) Monitoring variable values for the attack-free process with the proposed active detection method. (b) Monitoring variable values for the attack-free process with a time-triggered control system switching. . . . . . . . . . 93

4.3 (a) The output values over a simulation of the attack-free closed-loop process with the proposed active detection method. (b) The residual values over a simulation of the attack-free closed-loop process with the proposed active detection method. . . . . . . . . . . . . . . . . . . . . . . . . . . . 100

4.4 (a) The output values over a simulation of the attacked closed-loop process with the proposed active detection method. (b) The residual values over a simulation of the attacked closed-loop process with the proposed active detection method. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 101

4.5 (a) The attack detection times for different attack magnitudes and a dwell-time of $T_c = 1\,\mathrm{h}$. (b) The number of attacks detected under the attack-sensitive mode for an attack of magnitude of $\Lambda = \mathrm{diag}(1, 0.9)$ with different dwell-times. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 103

5.1 A block diagram illustrating a process control system under a false data injection attack that simultaneously alters the data over sensor-controller and controller-actuator communication links. . . . . . . . . . . . . . . . . 108

5.2 Illustrative example showing the reachable sets of the monitoring variable for the attack-free (blue sets) and the attacked (red sets) process in the presence of an undetectable attack, with two example trajectories (green lines) for the attacked process. . . . . . . . . . . . . . . . . . . . . . . . . 113

5.3 Illustrative example showing the reachable sets of the monitoring variable for the attack-free (blue sets) and the attacked (red sets) process in the presence of a detectable attack, with two example trajectories for the attacked process. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 114

5.4 (a) The state and estimation error values, (b) the monitoring variable values, used in the reachable set-based scheme, and (a)-(b) their corresponding reachable sets for the attack-free process over five time steps. . 118

5.5 (a) The reachable sets of the monitoring variable for the attack-free and the attacked process over a few time steps. At $k = 0$ and $k = 1$, $\mathcal{R}_k^{\eta} \cap \mathcal{R}_k^{\eta_a} = \emptyset$, indicating the attack is detectable. The localized zoom in the figure illustrates that the reachable sets at $k = 0$ do not intersect. (b) The values of the monitoring variable of the attacked process and the reachable sets used in the detection scheme over a few time steps. . . . . . . . . . . . 119

5.6 The reachable sets over a few time steps of the attack-free process and process under a multiplicative FDIA. Two initial sets are considered: (a) an initial set that is the attack-free minimum invariant set shifted by $\xi' = [10\ \ -5]^T$ and (b) an initial set that is the attack-free minimum invariant set shifted by $\xi' = [50\ \ -20]^T$. . . . . . . . . . . . . . . . . . . . . . . . . 121

5.7 The monitoring variable values and reachable sets used in the detection scheme over a few time steps for the process under a multiplicative FDIA. Two initial sets are considered: (a) an initial set that is the attack-free minimum invariant set by $\xi' = [10 \ \ -5]^T$ and (b) an initial set that is the attack-free minimum invariant set shifted by $\xi' = [50 \ \ -20]^T$. . . . . . . . 122

5.8 The reachable sets over a few time steps for the attack-free process with respect to the reachable sets for the process under (a) the first additive attack and (b) the second additive attack. . . . . . . . . . . . . . . . . . 123

5.9 The monitoring variable values and reachable sets used in the detection scheme over a few time steps for the attacked process. The monitoring variable values shown are observed over one simulation of the process under: (a) the first additive attack and (b) the second additive attack. . . . . . 124

5.10 The monitoring variable values, including (a) the output values and (b) the residual values, and reachable sets used in the detection scheme over a few time steps for the attack-free process. In this case, there are no false alarms. The brown central region represents the intersection of all reachable sets shown. . . . . . . . . . . . . . . . . . . . . . . . . . . 128

5.11 The monitoring variable values, including (a) the output values and (b) the residual values, and reachable sets used in the detection scheme over a few time steps for the attacked process. In this case, the attack is detected at $k = 1$. The brown central region represents the intersection of all reachable sets shown. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 130

5.12 The reachable sets for (a) the measured output and (b) the residual for the CSTR under an attack. . . . . . . . . . . . . . . . . . . . . . . 131

5.13 The values of (a) the measured output and (b) the residual and their corresponding terminal sets over one simulation of the CSTR process under an attack. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 132

6.1 The reachable sets of the attacked and the attack-free process under the nominal mode (a) during transient operation when the attack is potentially detectable, and (b) at the steady-state when the attack is undetectable with $\mathcal{R}^{\eta_a}_{\infty}(\mathcal{R}^{\xi_a}_{\infty}(K^N, L^N)') \subset \mathcal{R}^{\eta}_{\infty}(\mathcal{R}^{\xi}_{\infty}(K^N, L^N)')$. . . . . . . . . . . . . . . 155

6.2 Evolution of the monitoring variable values, with respect to the attack-free reachable sets, over a few time steps of a simulation of the scalar process with a control mode switch implemented at time step $t_{s_1} = 14$ for the case when (a) no attack takes place (demonstrating zero false alarms), and (b) an attack takes place at $t = 0$ and is detected at time step $t_d = 124$. . . 158

6.3 Evolution of the monitoring variable values observed over a few time steps of (a) the attack-free process with consecutive switches from the nominal to attack-sensitive mode implemented at $t_{s_1} = 221$, $t_{s_3} = 767$, and $t_{s_5} = 1313$ showing no false alarms due to switching and, (b) the attacked process with control mode switch from the nominal to the attack-sensitive mode implemented at time step $t_{s_1} = 221$ leading to attack detection at time step $t_d = 235$. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 163

6.4 Scheduled control mode switching for the CSTR process. A value of 0 on the Y-axis indicates that the process is operated under the nominal mode, and a value of 1 on the Y-axis indicates that the process is operated under the attack-sensitive mode. . . . . . . . . . . . . . . . . . . . . . . . . . 168

6.5 The attack schedule for the smart attack, which is designed to evade detection under the scheduled control mode switching-based detection scheme. The figures illustrate the attack schedule for: (a) the sensor-controller attack, and (b) the controller-actuator attack. . . . . . . . . . . . . . . . . 169

6.6 Values of the monitoring variable for the CSTR process under attack showing (a) no attack detection with scheduled control mode switching (b) attack detection at $t_d = 235$ under the randomized switching scheme after a control switch implemented at $t_{s_1} = 221$. . . . . . . . . . . . . . . . . . . 170

6.7 The values of the monitoring variable for the nonlinear CSTR under a smart attack showing: (a) no attack detection with a scheduled control mode switch, and (b) attack detection at time step $t_d = 231$ (2.31 h in real time) under the randomized switching scheme after a control mode switch at time step $t_{s_1} = 221$ (2.21 h in real time). . . . . . . . . . . . . . . . . . 172

7.1 Terminal sets of the monitoring variable for the attacked and the attack-free processes at the initial steady-state $x_i^s = 0$ with control parameters $(K^i, L^i)$. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 186

7.2 Reachable sets of the monitoring variable at $t = 1$ and the terminal sets of the monitoring variable for the attacked and the attack-free process under a control mode with $x_f^s = -2$ and $(K^f, L^f) = (1.5, 0.1)$. . . . . . . . . . . 188

7.3 Reachable sets of the monitoring variable for the attacked and the attack-free process at $t_d = 1$ under a control mode with $x_f^s = -30$ and $(K^f, L^f) = (1.5, 0.1)$. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 189

# List of Tables

2.1 Model parameters for the CSTR Process [1]. . . . . . . . . . . . . . . . . . . . 34

# Abstract

## Control-Enabled Approaches for Active Detection of Cyberattacks on Process Control Systems

Increasing reliance on wireless communication and complexity of cyberattacks have rendered industrial control systems (ICSs) such as process control systems (PCSs) (which are ICSs that operate chemical manufacturing processes) vulnerable to cyberattacks by malicious agents. In the past decade, several highly sophisticated cyberattacks (e.g., Stuxnet virus (2010), German steel mill attack (2014), Ukrainian power grid attack (2015), TRITON (2017)) have demonstrated that information technology (IT) infrastructure-based solutions to handling cyberattacks on control systems are insufficient on their own. An increasing body of research has focused on developing operational technology (OT)-based approaches to enhance the cyberattack resilience of PCSs. Cyberattack resilience here is defined as the ability of a PCS to minimize the impact of a cyberattack and recover from it. Research on cyberattack resilience of PCSs involves approaches that range from designing PCSs that are inherently attack-resilient to developing cyberattack detection, identification and mitigation schemes. Cyberattack detection schemes are OT-based anomaly detection schemes that reveal the presence of a cyberattack on a PCS by monitoring the process operational data for anomalies and are an important component of a cyberattack resilient PCS.

The motivating realization behind the work presented in this dissertation is that the influence of PCS design parameters may be exploited to reveal the presence of an ongoing cyberattack on a PCS. In the chapters that follow, several approaches for cyberattack detection are presented. First, a control screening approach that may be used to incorporate attack detectability within the conventional PCS design considerations is presented. The screening algorithm is based on a characterization of the interdependence between the PCS design parameters, and the ability of the detection scheme to detect the attack (attack detectability). Next, for a certain class of detection schemes monitoring a process, the relationship between the PCS design parameters, the closed-loop stability of the attacked process, and the detectability of certain attacks is rigorously characterized. Based on the characterization, for attack detection, it may be preferred to operate the process under performance degrading “attack-sensitive” parameters. To manage a potential tradeoff between attack detection and closed-loop performance, an active detection method utilizing switching between two control modes is developed. Under the active detection method, extended process operation is under a first (nominal) mode, the control parameters (called nominal parameters) for which are selected to meet traditional control design criteria. Under the second (attack-sensitive) mode, the process is operated with attack-sensitive parameters. The process is operated under the attack-sensitive mode intermittently to probe the process for an ongoing attack. Control parameter switching on a process under steady-state operation may induce transient behavior, which may trigger false alarms in the class of detection schemes. For processes with an invertible output matrix, a switching condition is imposed to select control parameter switching instances such that false alarms in the system are minimized.

To eliminate false alarms due to control switching on processes with a non-invertible output matrix, a reachable set-based detection scheme is developed. The reachable set-based cyberattack detection scheme guarantees a zero false alarm rate during transient attack-free process operation by tracking the evolution of the monitoring variable values with respect to their reachable sets of the attack-free process at each time step. Following this, a switching-enabled active detection method that utilizes the reachable set-based detection scheme to enable attack detection with a zero false alarm rate is presented. Furthermore, the control parameter switching instances between the nominal to attack-sensitive modes are randomized, thereby preserving the confidentiality of the detection method. Destabilization of a process for attack detection (as with operation under attack-sensitive mode) may not always be preferred. Two different alternate control modes that may be used to induce perturbations for active attack detection without destabilizing the attacked process are presented. To guarantee attack detection, the alternate control mode selected must induce “attack-revealing” perturbations in the process. Reachability analysis is used to present a set-based condition that if satisfied means that the control mode selected induces attack-revealing perturbations. Different models of false data injection attacks are considered. A screening algorithm that may be used to select an attack-revealing control mode for the active detection of attacks is presented. The application of all methods are applied to simulations of different illustrative processes to demonstrate their attack detection capabilities.

# Acknowledgments

First, I would like to express my profound gratitude to my advisors, Dr. Ellis and Dr. El-Farra for their guidance and mentorship. In addition to learning to write academic articles and presenting research findings succinctly, from Dr. Ellis, I learned how to be inquisitive and ensure the technical rigor of research by thoroughly questioning theses that may at first glance seem intuitive, a skill I will hone. From Dr. El-Farra, I learned how to perform research with curiosity, patience, and good spirit. I will be forever in awe of the breadth of Dr. El-Farra's technical knowledge, his ability to explain extremely difficult concepts with immense clarity, his patience, kindness, empathy, and humor. As I work towards being a mentor myself, I hope I am able to emulate both my advisors.

I would like to thank Dr. Palazoglu for serving on my dissertation committee, for chairing my qualifying examination committee, and for spending time giving me feedback on everything from presentations to planning an academic career. I am very grateful to Dr. Curtis and Dr. Manikantan, for all the time they spent guiding me as I work towards a career in academics. I am especially grateful to Dr. Curtis for giving me the opportunity to serve as her teaching assistant.

To my labmates, Antonea, Yue, Loren, Pranav, Aatam, Rahul, Hossein, and Sui. It was great to work with you. To Rahul, thank you for being great friend.

I would also like to thank my current and prior graduate coordinators Ryan Gorsiski and Grace Woods for being the most reliable and supportive people I have had the good fortune to have worked with.

Finally, I would like to thank my family and friends who have stood by me through the ups and downs of my doctoral journey.

# Chapter 1

# Introduction

## 1.1 Background and Motivation

In this section, the motivation for control-enabled approaches for ensuring cyberattack resilience of control systems is discussed. Following this, a review of the literature focused on the taxonomy of cyberattacks on process control systems (PCSs) and approaches for control-enabled approaches to PCS cyberattack resilience is presented. Finally, a brief overview of the work presented in this dissertation and the organization of the dissertation is presented.

### 1.1.1 Cyberattacks on Critical Infrastructure

The government of the United States of America has identified 16 infrastructure sectors as critical because their safe functioning is of vital importance to the security, economic security, national public health or safety, or any combination thereof of the United States of America [2]. An important sector among the critical infrastructure sectors is the chemical sector, as it manufactures, stores, uses, and transports potentially dangerous chemicals on which other critical infrastructure sectors rely [3]. Critical infrastructure sectors (including the chemical sector) rely on industrial control systems (ICSs) for their safe and economic operation. Increasing reliance on wireless communication and complexity of cyberattacks have rendered industrial control systems vulnerable to cyberattacks by malicious agents [4–6]. PCSs are ICSs that operate chemical manufacturing processes and may utilize networked communication to integrate the cyber components (e.g., controllers, human machine interfaces) with the physical components (e.g., sensors, actuators). Cyberattacks on PCSs may involve malicious incursions into the PCS network with the objective of altering the operational data communicated over the sensor-controller and/or controller-actuator communication links, and have increased in frequency over the past decade [7]. Cyberattacks resulting in a loss of control in the PCS operating a chemical manufacturing process may have disastrous consequences, and therefore, research focused on the enhancement of PCS cybersecurity has received increasing attention [8].

Traditionally, cyberattack threats to ICSs have been viewed as an information technology (IT) issue and addressed through IT infrastructure-based approaches [9]. However, in the past decade, several highly sophisticated cyberattacks (e.g., Stuxnet virus (2010), German steel mill attack (2014), Ukrainian power grid attack (2015), TRITON (2017) [10]) have demonstrated that IT infrastructure-based solutions for handling cyberattacks on control systems are insufficient on their own and must be augmented by operational technology (OT)-based approaches. OT-based approaches for enhancing cyberattack resilience may consider process operational data (e.g., measured variables, and control setpoints). Considerations for the design of an OT-based approach to enhance cyberattack resilience of a PCS may include a model of the attack. Attack models considered for the development of some approaches to cyberattack resilience of PCS in the literature are based on an elaborately explored taxonomy of cyberattacks based on several criteria, e.g., vulnerability exploited by the cyberattack and method that the cyberattack alters the PCS operation [11–18]. False data injection cyberattacks are the focus of the work presented in this dissertation and are discussed in the next section.

### 1.1.2 False Data Injection Cyberattacks

False data injection (FDI) attacks aim to compromise the integrity of the PCS by maliciously altering the data over the communication channels or within the PCS itself [18, 19]. The objective of an FDI attack could be to cause instability or adverse economic, environmental, or human life impacts. In designing an FDI attack, the major goal of attackers targeting process control systems (PCSs) is to falsify the process operational data by: (1) manipulating the data communicated over the controller-actuator communication link (controller-actuator attack), (2) manipulating the data communicated over the sensor-controller communication link (sensor-controller attack), or (3) manipulating the control system logic (Fig. 1.1).

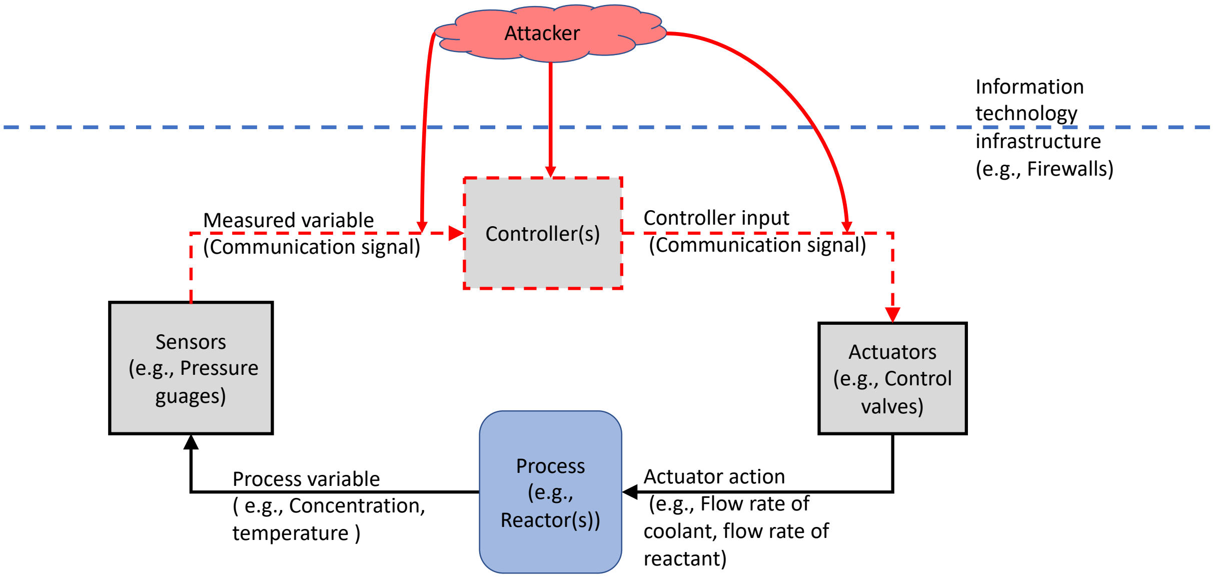

Attacker
Information technology infrastructure (e.g., Firewalls)
Measured variable (Communication signal)
Controller(s)
Controller input (Communication signal)
Sensors (e.g., Pressure guages)
Actuators (e.g., Control valves)
Process (e.g., Reactor(s))
Process variable ( e.g., Concentration, temperature )
Actuator action (e.g., Flow rate of coolant, flow rate of reactant)

Fig. 1.1: Block diagram of a PCS illustrating the potential targets for a false data injection cyberattack.

Chemical processes may be modeled by dynamics that are inherently complex (e.g., nonlinear, networked, spatially distributed). Consequently, an attacker may need some process knowledge to design of false-data injection attacks to cause instability or cause adverse economic, environmental, or human life impacts [20]. If an attacker has sufficient knowledge of the process, they may be able to design an attack that remains stealthy. Stealthy attacks may be defined as attacks that are designed to evade detection by falsifying process operational data in the PCS communication links by injecting data that is difficult to distinguish from the data of the attack-free process. Two main types of FDI attacks have received attention in the literature: (1) additive FDI attacks [21], and (2) multiplicative FDI attacks [22, 23]. Additive FDI attacks inject false data by adding a factor to the sensor measurement data communicated over the link, leading to the controller receiving the sensor measurement plus a factor added to it [21]. Additive FDI attacks may need careful design to remain stealthy, requiring process information (see Remark 5 in [22]). Multiplicative FDI attacks alter process operational data by multiplying the data over the communication link by a factor [23], and are unique because they may be designed to be stealthy without requiring intimate knowledge of process dynamics (see Remark 6 in [22]). This dissertation considers the detection of FDI attacks that alter data over one or both of the sensor-controller and the controller-actuator communication links. FDI attacks involving multiplicative, additive, and mixed multiplicative and additive attacks are studied. In the next section, the notion of cyberattack resilience considered in this work and some approaches proposed in the literature for control-enabled cyberattack resilience are discussed.

### 1.1.3 Control-Enabled Approaches for Cyberattack Resilience

In the literature, several approaches considering the design of OT-based approaches for ensuring cyberattack resilience of PCS have been proposed [21, 23–40]. Cyberattack resilience here is defined as the ability of a PCS to minimize the impact of a cyberattack and recover from it. Due to the interconnected nature of a cyber-physical system like a PCS, a comprehensive solution to enhancing the cybersecurity of PCSs may involve adopting a multi-faceted approach that involves both information technology (IT)-based and operational technology(OT)-based approaches. Approaches proposed in the literature broadly include those that reinforce IT-based infrastructure (e.g., firewalls [41]), approaches that involve cybersecure architecture design which may utilize redundant sensors and actuators as a backup (e.g. [26]), approaches for process equipment design to mitigate adverse impacts of a cyberattack (e.g. [42, 43]), and the design of OT-based approaches for detecting, identifying, and successfully recovering from an ongoing cyberattack [22, 26, 44–54].

An approach for design of a control law for deterring additive FDI attacks involves characterizing ellipsoids that are outer estimates of the set of estimation errors (termed hidden reachable sets) induced by a cyberattack using linear matrix inequalities (LMIs) [39], and leveraging them for an LMI-based controller design methodology that limits the reachable sets under attack while ensuring that certain performance considerations are met[25]. As a result of its design, this controller restricts the magnitude of a successful attack on the process, making it inherently attack resilient. Another approach explored, involves implementing a randomized control law that changes at periodic intervals of time while ensuring its closed-loop stability [24]. This control law design prevents an attacker from understanding the control law based on the process operational data in the communication links of the PCS, and as a consequence, prevent the attacker from designing an attack on it. These methods consider the design of a control law that makes a PCS inherently cyberattack resilient. The focus of this dissertation, however, is the design of cyberattack detection approaches.

Cyberattack detection schemes monitor the data in the PCS communication links using a monitoring metric to detect an anomalous behavior in the PCS. Their design has received extensive attention in the literature [26, 27, 31–33]. One approach to designing a detection scheme in the literature involves using a neural network-based control detection strategy and a cybersecure PCS architecture design with built-in redundancies [26]. The requirement of all possible attack scenarios for training may make this approach difficult to implement practically. Another approach to attack detection considers nonlinear systems under Lyapunov-based model predictive control with data-driven models and presents a detection scheme that detects an attack based on when the data-driven model becomes insufficiently accurate for maintaining the closed-loop state of the process within a desired region of state-space [27]. Attack identification and mitigation schemes aim to identify the magnitude of a cyberattack, and mitigate its impact on a PCS. Some approaches explored in the literature for designing these schemes combine identification and mitigation and present novel observer designs, which are able to identify attacked sensors, and discard falsified measurements from control law computation for a PCS [28–30, 35]. Other approaches focus on cyberattack mitigation after an attack detection, and involve control reconfiguration upon the detection/identification of an attack [34] or varying an external signal and achieving adaptive stabilization in the presence of an ongoing cyberattack [36]. In the next section, some cyberattack detection schemes proposed in the literature are discussed.

### 1.1.4 Detection of Cyberattacks on PCSs

An important component of the cyberattack resiliency of PCSs is the ability to detect the presence of a cyberattack by a cyberattack detection scheme. For PCS cyberattack detection, many cyberattack detection schemes have been proposed [23, 26, 55–64]. Attack detection methods can be broadly divided into two categories, including passive detection schemes and active attack detection methods. Passive attack detection schemes monitor a process for anomalies based on regular operational data without employing external intervention or applying a perturbation. These schemes have been extensively explored [21, 38, 65, 66]. For example, one passive scheme differentiates the behavior of an attacked process from its attack-free behavior by characterizing the skewness in the detection metric distribution [65]. Another approach uses a two-tier controller-detector architecture, with a neural network-based detection scheme to monitor for some attacks [26]. Other approaches use standard residual-based detection schemes such as the cumulative sum (CUSUM) or $\chi^2$ detection schemes to identify anomalous behavior [21, 38, 66]. FDI attacks targeting phasor measurement units have been considered where conditions were derived for undetectable additive and multiplicative attacks with a standard detection scheme (e.g., $\chi^2$ detection scheme) [66]. An enhanced detection scheme was proposed. Closed-loop systems, where falsified output measurements are used in the controller, were not considered. The use of the CUSUM and $\chi^2$ detection schemes for monitoring closed-loop systems under additive false-data injection sensor-control link attacks was considered in [21, 38].

Passive approaches for attack detection may not always be successful in differentiating the anomalous behavior in the attacked process from its attack-free behavior (e.g., Section III, [62]). As an alternative, an active detection method may potentially enhance the detection capabilities. Active attack detection methods involve external intervention to induce an attack-detecting perturbation in the closed-loop process [23, 62–64, 67]. Two active detection methods were presented in [62]. The first approach utilizes a watermarking scheme, i.e., a secret noisy input is added to the computed control input to the process, and an attack is detected if the distribution of the detection metric deviates from the distribution expected for an attack-free process. The second approach uses a moving target scheme, i.e., the original system is augmented with an authenticating subsystem with time-varying dynamics and additional sensors to estimate the subsystem state. In this approach, the attack detection scheme is based on the difference between the (potentially) falsified output and the expected output. An approach that uses a combination of watermarking and a moving target scheme has also been explored [63]. In [67], the detectability of stealthy attacks exciting the zero-dynamics was characterized as a function of the observability of the attacked process. Leveraging this characterization, detection schemes that use redundant sensors and actuators were proposed to enable the detection of a zero-dynamics exciting attack. A few active detection approaches for the detection of multiplicative cyberattacks have been proposed [23, 64]. Specifically, a watermarking approach that adds a constant to the sensor measurement before communicating the resulting value to the controller was proposed for multiplicative sensor-controller link attacks [23]. The controller subtracts the constant before computing its control action. Another watermarking scheme was presented in [64], utilizing an additive secret signal with a known distribution to the control input to detect several attacks such as multiplicative cyberattacks. The sensor data reported by all sensors are subject to two tests developed based on the statistical hypothesis testing criterion.

While detectability may be thought of as a systems-theoretic property, the detectability of an attack, defined herein as the ability of a detection scheme to detect an attack, may depend on the class of detection schemes used to monitor the process. PCS design parameters, and the mode of operation of the process (e.g., steady-state or transient) may influence the detectability of an attack. Prior to the work discussed in the succeeding chapters, this dependence between the PCS parameters, process operation, and the detectability of an attack was unexplored in the literature.

## 1.2 Objectives and Organization of the Dissertation

The main objective of this dissertation is to develop methods that may be used to detect a cyberattack on a PCS by exploiting the interdependence of PCS design parameters (e.g., controller gain, observer gain), the stability of the closed-loop system, the mode of operation (e.g., steady-state or transient) of the process, and the ability of a detection scheme to detect attacks (defined as attack detectability) with respect to certain classes of cyberattack detection schemes.

Broadly, the dissertation is organized as follows:

- In Chapters 2-4, the detection of multiplicative FDI attacks that alter the data over the sensor-controller communication channels of a PCS that operates a process under steady-state conditions is considered. Passive and active detection methods are presented.

- In Chapter 5, attacks that may simultaneously alter the data over the sensor-controller and controller-actuator communication channels are considered. The attacks considered may be classified into attacks that can be modeled as: (1) only multiplicative FDI attacks, (2) only additive FDI attacks, and (3) a combination of additive and multiplicative attacks. A passive detection scheme that generates no false alarms when monitoring the process under transient operation is presented.

- In Chapter 6, multiplicative FDI attacks that may simultaneously alter the data over the sensor-controller and controller-actuator communication links are considered. A randomized control mode switching-enabled active detection method that guarantees attack detection with no false alarms, while preserving the confidentiality of the detection method, is presented.

- In Chapter 7, FDI attacks that may be modeled as a combination of multiplicative and additive attacks, and those that simultaneously alter the data over the sensor-controller and controller-actuator communication links are considered. An approach for control mode selection for an active detection approach that guarantees attack detection without destabilizing the attacked process is presented.

- In Chapter 8, conclusions of the work presented in this dissertation and a brief discussion about potential directions for future research are presented.

The chapters are organized as follows. Chapter 2 presents an approach for enhancing PCS cyberattack resiliency by incorporating the detectability of an attack as a criterion into conventional control design criteria (e.g., closed-loop stability and economic considerations). Specifically, a controller screening methodology aimed at identifying controller parameter choices that mask the detectability of a range of multiplicative FDI attacks with respect to a class of detection schemes is presented. The work presented in this chapter is based on a version of the published paper [22].

In Chapter 3, the relationship between closed-loop stability, PCS parameters (controller and observer gain), and attack detectability with respect to a class of attack detection schemes is rigorously characterized. The results are used to identify PCS parameters (called “attack-sensitive” parameters) such that an attack on the process operated under attack-sensitive parameters destabilize the process, and enable attack detection. The selection of attack-sensitive control system parameters can enhance the ability to detect attacks, but can also degrade the performance of the attack-free process. An active attack detection methodology employing control system parameter switching is developed to manage the performance degradation in the attack-free process operated with attack-sensitive parameters. Under the proposed active detection method, the extended operation of the process is under “nominal” control parameters that are chosen to meet standard controller design criteria. To probe for attacks, the process is intermittently operated with attack-sensitive parameters. The work presented in this chapter is based on a version of the published papers [50, 68].

The active detection methodology, presented in Chapter 3, is developed for processes under steady-state operation, when the process states remain bounded within a small neighborhood of the desired operating steady-state. Control parameter switching on the process under steady-state operation may excite process dynamics, and result in a brief transient operation of the process, when its states are not bounded within a small neighborhood of the steady-state. Transient process operation may generate false alarms in the detection scheme. To this end, false alarm minimization under control switching is considered in Chapter 4, where a control switching approach that enables attack detection with minimal false alarms is presented. Processes with an invertible output matrix are considered. A state-dependent switching condition that guarantees zero false alarms is developed using a region that is likely to contain the attack-free process states, called the confidence region. Practical implementation issues related to the active detection method are discussed, including the inability to ensure that the switching condition will be satisfied over the time interval it is desired to switch the control system. To minimize false alarms in the event that a control parameter switch is implemented when the switching condition is not satisfied, an alarm suppression scheme is discussed. The work presented in this chapter is based on a version of the published paper [48].

The class of detection schemes considered in Chapter 3 are tuned for attack-free process operation under steady-state conditions, when the process states may be bounded within a small region containing the steady-state. Another approach for eliminating false alarms in a detection scheme due to control parameter switching may be to utilize a detection scheme tuned for transient process operation. In Chapter 5, a reachable set-based detection scheme is proposed to monitor transient process operations. FDI attacks that alter the variable value communicated over both the sensor-controller and controller-actuator PCS communication links, and those that may be modeled as additive or multiplicative or as both additive and multiplicative attacks, are considered. The proposed detection scheme verifies whether the value of a generalized monitoring variable at a given time step is contained within its reachable set for the attack-free process. The proposed detection scheme monitors the process without requiring extensive closed-loop data. It also does not raise false alarms during transient operation, without placing any assumptions on the structure of the output matrix. Conditions that lead to an attack being detectable or undetectable with respect to the proposed detection scheme are characterized. The work presented in this chapter is based on a version of the published papers [69, 70].

In Chapter 6, multiplicative attacks that simultaneously alter the data over the sensor-controller and controller-actuator communication links are considered. A switching-enabled active detection method that enables attack detection with a guaranteed zero false alarms from a control parameter switch is presented. Theoretical results are presented to demonstrate that the randomized switching-enabled attack detection method guarantees a zero false alarm rate from multiple successive control mode switches on the process (under transient operation), without requiring that the output matrix be invertible. Some implementation considerations are discussed, and algorithms for two variations of the switching-enabled attack detection method are presented. Randomization of control parameter switching instances is considered as a way to preserve the confidentiality of the detection method. The work presented in this chapter is based on a version of the published paper [71].

The proposed method in Chapter 6 enables attack detection by selecting attack-sensitive control parameters such that an attack destabilizes the process operated with attack-sensitive parameters. Destabilization for attack detection may not always be preferred. In Chapter 7, attack detection without destabilizing the attacked process is considered. An approach for selection of alternative active detection method(s) for detecting false-data injection cyberattacks that alter the data communicated over the PCS communication channels is presented. In particular, two alternative control modes, one involving changing set points and the other involving switching control parameters, are considered for the active detection of a class of stealthy false data injection attacks. Implementing either control mode induces perturbations in the closed-loop process. To guarantee the detection of an attack, the perturbations induced from implementing a control mode on the attacked process should be "attack-revealing." Reachability analysis is used to present a condition that if satisfied means that an attack will be detected, forming the basis of attack-revealing perturbations. Using the condition, a screening algorithm that may be used to choose a control mode that guarantees the detection of an attack is presented. The work presented in this chapter is based on a version of the published paper [72]. At the end of each chapter, the application of the approaches discussed in that chapter is demonstrated utilizing simulations of illustrative process(es). The proofs for all propositions are presented in the appendices.

# Chapter 2

# Detectability-Based Controller Design Screening for Processes Under Multiplicative Cyberattacks

In this chapter, a screening methodology that can be used during the PCS design phase to identify and discard control parameters that mask multiplicative attacks on the sensor-controller communication link from being detected is presented. The methodology may be used to incorporate cyberattack detection as a criterion for controller design in addition to the criteria commonly used in practice (e.g., closed-loop stability, process economics, and robustness to uncertainty [73–75]). Processes that can be modeled by discrete-time linear time-invariant (LTI) systems subject to bounded measurement noise and process disturbances are considered. To characterize the detectability of an attack, a general class of residual-based process monitoring detection schemes is considered. A residual set-based condition is developed, which is a function of the controller gain, the observer gain, and attack magnitude, and the condition is used to characterize an attack as potentially detectable or undetectable with respect to the class of residual-based detection schemes considered. The condition provides an explicit connection between the PCS design and detectability of an attack. Subsequently, the residual set-based condition is incorporated into the proposed screening methodology. The application of the screening methodology is demonstrated using two illustrative examples.

## 2.1 Preliminaries

### 2.1.1 Notation and Definitions

The $n$-dimensional Euclidean space is denoted by $\mathbb{R}^n$. Closed 2-norm and $\infty$-norm balls with radius $\epsilon > 0$ are denoted as $B^n(\epsilon)$ and $B^n_\infty(\epsilon)$, respectively, where $B^n(\epsilon) := \{\xi \in \mathbb{R}^n \mid \|\xi\| \leq \epsilon\}$ and $B^n_\infty(\epsilon) := \{\xi \in \mathbb{R}^n \mid \|\xi\|_\infty \leq \epsilon\}$. The Minkowski sum of the sets $U \subset \mathbb{R}^n$ and $V \subset \mathbb{R}^n$ is denoted by $U \oplus V = \{u + v \mid u \in U, v \in V\}$. The Minkowski difference of the sets $U \subset \mathbb{R}^n$ and $V \subset \mathbb{R}^n$ is denoted by $U \ominus V = \{u - v \mid u \in U, v \in V\}$. The matrices $I_{n\times n}$ and $0_{n\times m}$ denote the $n \times n$ identity matrix and the $n \times m$ zero matrix, respectively. $\lambda_i(A)$ is the $i^{th}$ eigenvalue of the matrix $A$. The support function of a nonempty set $X \subset \mathbb{R}^n$ evaluated at $a \in \mathbb{R}^n$ is defined as $h_X(a) = \sup_{x\in X} a^T x$. A polytope is a bounded polyhedron and is the solution set of a finite number of linear inequalities ($P = \{x \in \mathbb{R}^n \mid {a_i}^T x \leq b_i, i = 1, 2, \ldots, M\}$ where $M$ is the number of faces of the polytope $P$) [76].

### 2.1.2 Process Model and Control System

Processes modeled by discrete-time LTI systems are considered:

$$x(t + 1) = Ax(t) + Bu(t) + Gw(t) \tag{2.1}$$

where $x(t) \in \mathbb{R}^n$ is the state vector, $u(t) \in \mathbb{R}^p$ is the manipulated input vector, $w(t) \in W \subset \mathbb{R}^{n_w}$ is the process disturbance vector, and $A$, $B$ and $G$ are matrices of appropriate dimensions. The set $W$ is a compact set that describes the admissible process disturbances. In this work, multiplicative cyberattacks that manipulate data over the sensor-controller communication link are considered and are modeled by:

$$y(t) = \Lambda(Cx(t) + v(t)) \tag{2.2}$$

where $y(t) \in \mathbb{R}^m$ is the output vector, $v(t) \in V \subset \mathbb{R}^m$ is bounded measurement noise, and $C \in \mathbb{R}^{m\times n}$ is the output matrix. Under attack-free conditions, the multiplicative attack matrix is $\Lambda = I_{m\times m}$, while in the presence of an attack, is $\Lambda = diag(\alpha_1, \alpha_2, \ldots, \alpha_m)$ where $\alpha_i \neq 1$ represents the magnitude of attack on the $i^{th}$ sensor. The matrix $\Lambda$ is referred to as the attack magnitude. The sets $W$ and $V$ are assumed to be polytopes that contain the origin.

A Luenberger observer is used to estimate the state as follows:

$$\hat{x}(t+1) = A\hat{x}(t) + Bu(t) + L(y(t) - \hat{y}(t)) \tag{2.3}$$

where $\hat{x}(t) \in \mathbb{R}^n$ represents the estimated state, $\hat{y}(t) = C\hat{x}(t)$ represents the estimated output of the process, and $L \in \mathbb{R}^{n\times m}$ is the observer gain selected so the eigenvalues of $A - LC$ are within the unit circle. The desired steady-state is taken to be origin. To steer the state to the origin, a linear feedback control law of the following form is used:

$$u(t) = -K\hat{x}(t) \tag{2.4}$$

where $K \in \mathbb{R}^{p\times n}$ is the controller gain selected such that the eigenvalues of $A - BK$ are within the unit circle.

The estimation error is defined as $e(t) = x(t) - \hat{x}(t)$ and evolves according to:

$$e(t+1) = L(I - \Lambda)Cx(t) + (A - LC)e(t) + Gw(t) - L\Lambda v(t) \tag{2.5}$$

Defining the augmented state vector as $\xi(t) = [x^T(t)\ e^T(t)]^T$, the overall closed-loop process consisting of the process (Eq. 2.1) with the feedback control law (Eq. 2.4) using the estimated state from the observer (Eq. 2.3) can be written as:

$$\xi(t+1) = \begin{bmatrix} (A-BK) & BK \\ L(I-\Lambda)C & (A-LC) \end{bmatrix} \xi(t) + \begin{bmatrix} G & 0_{n\times n} \\ G & -L\Lambda \end{bmatrix} d(t) = A_\xi \xi(t) + B_\xi d(t) \tag{2.6}$$

where $d(t) = \begin{bmatrix} w^T(t) & v^T(t) \end{bmatrix}^T \in F$ is the disturbance and measurement noise and $F = \left\{ \begin{bmatrix} w \\ v \end{bmatrix} \mid w \in W, v \in V \right\}$. For clarity of presentation, $\xi(t)$ denotes the augmented state in the absence of an attack, $\xi_a(t)$ denotes the augmented state in the presence of an attack, $A_\xi$ and $B_\xi$ are the system matrices for the augmented attack-free state dynamics ($\Lambda = I$), and $A_{\xi_a}$ and $B_{\xi_a}$ are the system matrices for the augmented state dynamics in the presence of an attack ($\Lambda \neq I$).

Owing to the persistent perturbation $d(t)$, closed-loop stability of the augmented closed-loop process (Eq. 2.6) is characterized by ultimate boundedness of the augmented state. When the eigenvalues of $A_\xi$ (or $A_{\xi_a}$ in the presence of an attack) are within the unit circle, the augmented state converges to a terminal set containing the origin where it is maintained thereafter. The origin is denoted by $\xi_s^T := \begin{bmatrix} 0^T & 0^T \end{bmatrix}$. The terminal set is the minimum invariant set of the augmented closed-loop system [77]. The minimum invariant set denoted by $D_\xi$, is the Minkowski sum of the infinite series [77]:

$$D_\xi = B_\xi F \oplus A_\xi B_\xi F \oplus {A_\xi}^2 B_\xi F \oplus \dots \tag{2.7}$$

Similarly, for the attacked process, the minimum invariant set denoted by $D_{\xi_a}$ is given by:

$$D_{\xi_a} = B_{\xi_a} F \oplus A_{\xi_a} B_{\xi_a} F \oplus {A_{\xi_a}}^2 B_{\xi_a} F \oplus \dots \tag{2.8}$$

### 2.1.3 Residual-Based Attack Detection Scheme

A general class of residual-based attack detection schemes employed in process monitoring to detect abnormal operation is considered. The residual vector is defined as:

$$r(t) = y(t) - \hat{y}(t)$$

The residual characterizes the deviation of the measured output from the expected output. Writing the residual in terms of the augmented state and $d(t)$ for the attack-free system for analysis purposes gives:

$$r(t) = \begin{bmatrix} 0_{n\times n} & C \end{bmatrix} \xi(t) + \begin{bmatrix} 0_{m\times n} & I \end{bmatrix} d(t) = A_r \xi(t) + B_r d(t) \tag{2.9}$$

and similarly, in the presence of an attack, gives:

$$r(t) = \begin{bmatrix} (\Lambda - I)C & C \end{bmatrix} \xi_a(t) + \begin{bmatrix} 0_{m\times n} & \Lambda \end{bmatrix} d(t) = A_{r_a} \xi_a(t) + B_{r_a} d(t) \tag{2.10}$$

A steady-state analysis is presented in this paper, i.e., the augmented states $\xi(t)$ and $\xi_a(t)$ are considered to be bounded in a minimum invariant set for all time $t \geq 0$. This analysis reflects the normal operating conditions of processes around their steady-state. Since the sets $D_\xi$ and $D_{\xi_a}$ are positively invariant, and $F$ is a polytope, the residual vector $r(t)$ is bounded within a compact terminal set $D_r = A_r D_\xi \oplus B_r F$ ($D_{r_a} = A_{r_a} D_{\xi_a} \oplus B_{r_a} F$), for all $\xi(t) \in D_\xi$ ($\xi_a(t) \in D_{\xi_a}$) and $d(t) \in F$.

The class of residual-based detection schemes considered are those that monitor the 2-norm of the residual ($\|r(t)\|$) to identify anomalous behavior and are defined as follows:

$$z_d(t) = f_d(z_d(t-1), \|r(t)\|; p); \;\; z_d(-1) = 0 \tag{2.11a}$$

$$z_s(t) = f_s(z_d(t), \|r(t)\|; p) \tag{2.11b}$$

where $z_d(t)$ represents the dynamic component of the detection scheme and $z_s(t)$ the output of the detection scheme. The parameterized functions $f_d(\cdot, \cdot; p)$ and $f_s(\cdot, \cdot; p)$ are continuous functions in both arguments for a given parameter $p$ where $p$ is a detector-specific tuning parameter. The function $f_s$ is a scalar-valued function.

The detection schemes are tuned with detector-specific parameter(s) and an alarm threshold ($\tau > 0$) chosen such that $z_s(t) \leq \tau$ during normal operation and $z_s(t) > \tau$ during abnormal operation. Specifically, the detection schemes are assumed to be tuned with all associated parameters chosen such that they generate zero alarms during normal operation. If $r(t) \in D_r$, the detector-specific parameters and the threshold $\tau$ are chosen such that $z_s(t)$ never breaches the threshold for all $t > 0$, i.e., $z_s(t) \leq \tau$. Additionally, if $\|r(t)\| > \max\limits_{r \in D_r} \|r\|$ for all time $t > 0$, then for sufficiently large $t$, $z_s(t) > \tau$ leading to a positive detection of anomalous process behavior. No assumption is made on residual sequences that take values both in $D_r$ and outside $D_r$ over time. These assumptions around tuning of the detection schemes are made to analyze the detection of attacks in a deterministic setting and to aid in establishing the definitions of detectable and undetectable attacks in the subsequent sections.

#### 2.1.3.1 Examples of Residual-Based Detectors

In practice, statistical process control schemes from the literature (e.g., $\chi^2$-squared, cumulative sum (CUSUM), and exponentially weighted moving average (EWMA) [78]) monitoring the 2-norm of residual ($\|r(t)\|$) may be considered as those fitting the class of residual-based detection schemes considered. A $\chi^2$-squared detection is a static detection

scheme tuned to generate an alarm when the value of the detection metric ($\|r(t)\|$) exceeds a pre-determined alarm threshold ($\tau_\chi$) [21]. This detection scheme may be formulated to fit the model introduced in Eq. 2.11 with $z_s(t) = \|r(t)\|$ and $\tau = \tau_\chi$. To tune the detection scheme to generate zero alarms in the absence of an attack, the threshold for the detection scheme may be chosen as $\tau_\chi = \max_{r \in D_r} \|r\|$.
The CUSUM detection scheme [37] is a sequential analysis hypothesis testing methodology, developed based on the sequential probability ratio test algorithm. Specifically, the CUSUM detection scheme is formulated as an integrator measuring the drift of the detection metric from a pre-determined baseline value of the metric given by:

$$S(t) = \max(S(t-1) + \|r(t)\| - b, 0);\ S(-1) = 0 \tag{2.12}$$

where $b$ is the baseline value of the metric and the detector-specific parameter. The scheme fits the scheme in Eq. 2.11 by letting $z_d(t) = z_s(t) = S(t)$, $\tau = \tau_s$ and $p = b$. To produce zero false alarms under normal operation, the detector is tuned with the choice of $b = \max\limits_{r \in D_r} \|r\|$ which ensures that, the CUSUM statistic always remains at zero in the absence of an attack ($S(t) = 0$ for all $r(t) \in D_r$). The threshold can be selected as any small positive number $\tau_C > 0$.

**Remark 2.1.1.** *Analyses and results presented herein do not make any assumptions on the stochastic distribution of the process disturbance and measurement noise, and approach the tuning of the detection scheme from a purely deterministic standpoint. To aid in the characterization of zero-alarm attack conditions, the tuning method adopted ensures that there are zero alarms during attack-free operation of the process. In practice, process monitoring schemes are often tuned to manage the trade-off between the false alarm rate during normal operation and the alarm rate during abnormal operation. As a result, while the analyses and results presented in subsequent sections may provide some insight into the detectability of attacks, they do not address the characterization of attack detectability with respect to detection schemes tuned based on other approaches. To characterize the false alarm rate and to characterize the probability of detecting an attack, the probability distributions of the process disturbance and the measurement noise must be available.*

*Here, the only assumption placed on the process disturbance and measurement noise is that they are bounded within a compact set. Characterizing the false alarm rate and probability of detection are beyond the scope of the work presented in this chapter.*

## 2.2 Controller Screening Methodology

In this section, multiplicative sensor-controller attacks on a process are classified as detectable, undetectable, and potentially detectable based on their impact on the class of residual-based detection schemes of the form Eq. 2.11. Then, based on the classification of an undetectable attack, a residual set-based undetectability condition is presented. If the condition is satisfied for a given multiplicative sensor-controller attack with magnitude $\Lambda$, controller gain $K$ and observer gain $L$, then the attack is undetectable. Following this, a numerically implementable approach for verifying the undetectability condition is presented. Finally, leveraging the undetectability condition, a controller screening methodology to identify and discard control parameters ($K$ and $L$) that mask the detectability of an attack of a predetermined magnitude is presented.

### 2.2.1 Detectability-Based Classification of Attacks

With respect to the class of residual-based detection schemes defined in Eq. 2.11, if the residuals of the process in Eq. 2.1-2.2 under an attack of magnitude $\Lambda \neq I$ are such that the output of the detection scheme $z_s(t) \leq \tau$ for all $t \geq 0$, zero alarms are raised and the attack is not detected. If the attack is not detectable in general, i.e., the detection scheme output is $z_s(t) \leq \tau$ for all $t \geq 0$ and for all $\xi(0) \in D_\xi$ under an attack, then the attack is defined as an undetectable attack. Similarly, if the residuals of the attacked process are such that they result in a breach of threshold by the output of the detection scheme $z_s(t) > \tau$ at some time $t \geq 0$, then the attack is called a detectable attack. If an attack has a nonzero probability of being detected, i.e., there exists a possible realization of the residual such that $z_s(t) \geq \tau$ for some $t$, then the attack is called a potentially detectable attack. Based on these definitions, this section presents conditions for undetectable, detectable, and potentially detectable attacks.

Any attack of magnitude $\Lambda \neq I$ on the process in Eq 2.1-2.2 can be defined as an undetectable attack with respect to the class of residual-based detection schemes considered in Eq. 2.11 if the residual under attack is such that $\|r(t)\| \leq R$ for all $t \geq 0$ where $R := \max_{r' \in D_r} \|r'\|$. Specifically, based on the tuning of the residual-based detection scheme as discussed in the previous section, if $\|r(t)\| \leq R$ for all $t \geq 0$, the output of the detection scheme never breaches the threshold. Thus, $z_s(t) \leq \tau$ for all $t \geq 0$, and the attack is undetectable. For the residual sequence to remain bounded for the closed-loop process under an attack, the closed-loop process must remain stable in the sense that the eigenvalues of $A_{\xi_a}$ must lie within the unit circle. Otherwise, $\xi_a(t)$ is unbounded implying that $r(t)$ is unbounded based on Eq. 2.10. Since $\|r(t)\| \leq \max_{r'_a \in D_{r_a}} \|r'_a\|$ for all $t \geq 0$ under a multiplicative attack, $\|r(t)\| \leq R$ is satisfied if:

$$R_a \leq R \tag{2.13}$$

where $R_a := \max_{r'_a \in D_{r_a}} \|r'_a\|$ and $D_{r_a} \subseteq B^n(R_a)$. This establishes a sufficient residual set-based condition that must be satisfied for an attack to be undetectable as summarized in the following proposition.

**Proposition 1.** *Consider the closed-loop process consisting of the process (Eq. 2.1) under the controller (Eq. 2.4) using the state estimate from the observer (Eq. 2.3) and subjected to a multiplicative sensor-controller link attack with magnitude $\Lambda \neq I$. Let the eigenvalues of $A_\xi$ and $A_{\xi_a}$ lie within the unit circle. For the class of residual-based detection schemes as defined in Eq. 2.11, a multiplicative sensor-controller link attack of magnitude $\Lambda \neq I$ on the process in Eq. 2.1-2.2 is undetectable if $R_a \leq R$.*

Since the disturbance set $F$ contains the origin, the sets $D_\xi$ and $D_{\xi_a}$ contain the origin. Also, the sets $D_r = A_r D_\xi \oplus B_r F$ and $D_{r_a} = A_{r_a} D_{\xi_a} \oplus B_{r_a} F$ contain the origin as they are Minkowski sums of linear transformations of the sets $D_\xi$, $D_{\xi_a}$ and $F$. As a result, for the process and attack model considered, the sets $D_r$ and $D_{r_a}$ are intersecting sets or $D_r \cap D_{r_a} \neq \emptyset$.

If the closed-loop process under attack is unstable, i.e., $\max|\lambda(A_{\xi_a})| > 1$, then the residual sequence is unbounded, and the output of the detection scheme will exceed the threshold $z_s(t) > \tau$ for a sufficiently large $t > 0$ resulting in a detection of the attack. Thus, all attacks that render the closed-loop process unstable are detectable attacks. If the closed-loop process under attack is stable but $\|r(t)\| > R$ for all $t \geq 0$, then the attack is detectable. However, no multiplicative sensor-controller attack that maintains closed-loop stability is such that $\|r(t)\| > R$ for all $t \geq 0$ since $D_{r_a} \cap D_r \neq \emptyset$. Possible realizations of the residual sequence exist such that $r(t) \in D_{r_a} \cap D_r$ for some $t$ which implies that $\|r(t)\| \leq R$ for some $t$. Therefore, when $R_a > R$, the attack may be potentially detectable with respect to the class of residual-based detectors considered.

The crucial observation from this analysis is that the controller and observer parameters $K$ and $L$ both play a role in whether a multiplicative attack is undetectable or potentially detectable since both sets $D_r$ and $D_{r_a}$ depend on $K$ and $L$ through the matrices representing the dynamics of the augmented state, $A_\xi$ and $B_\xi$ and $A_{\xi_a}$ and $B_{\xi_a}$, respectively. This observation forms the basis for incorporating cybersecurity considerations in the PCS design in this work. However, since $D_r$ and $D_{r_a}$ cannot be exactly computed in general, numerical estimates of $D_r$ and $D_{r_a}$ can be used to check for choices of $K$ and $L$ that mask the detectability of an attack of a given magnitude $\Lambda$ and discard them in favor of gains that do not mask its detectability.

A special approach for classifying a multiplicative sensor-controller attack as undetectable is when $r_a(t) \in D_r$ for all $t > 0$, i.e., the sets $D_r$ and $D_{r_a}$ are such that:

$$D_{r_a} \subseteq D_r \tag{2.14}$$

If the condition in Eq. 2.14 is satisfied, then, $\|r_a(t)\| \leq R$ for all time $t > 0$, leading to $z_s(t) < \tau$ and the attack is undetectable. This condition is a more restrictive condition for identifying undetectable attacks and does not account for all undetectable attacks because the residual-based detection schemes considered use the 2-norm of the residuals as a detection metric. As a result, the detection schemes considered do not account for the shape of the sets $D_r$ and $D_{r_a}$, which are not balls in general. Thus, if $D_{r_a} \subseteq D_r$, then $B^n(R_a) \subseteq B^n(R)$ and $R_a \leq R$. If Eq. 2.14 is satisfied, Eq. 2.13 is also satisfied. The converse is not true. Fig. 2.1 illustrates an example of a case with $R_a \leq R$ but $D_{r_a} \not\subseteq D_r$.

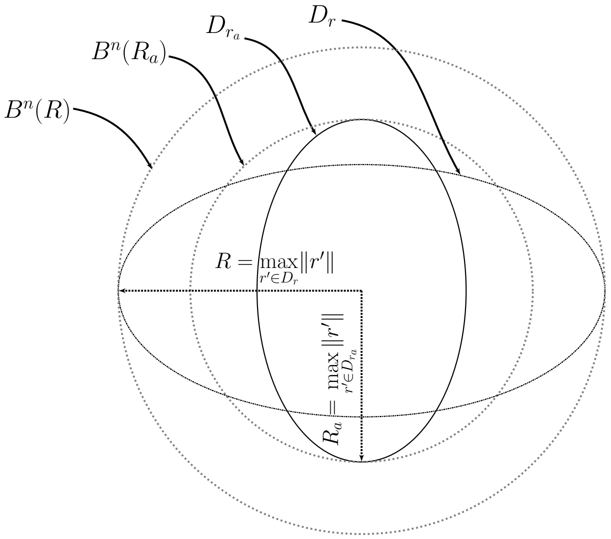

$D_r$
$D_{r_a}$
$B^n(R_a)$
$B^n(R)$
$R = \max_{r' \in D_r} \|r'\|$
$R_a = \max_{r' \in D_{r_a}} \|r'\|$

Fig. 2.1: Example of an undetectable attack with $R_a \leq R$ but $D_{r_a} \not\subset D_r$.

**Remark 2.2.1.** *Since the class of residual-based detection schemes considered in Eq. 2.11 monitor the process based using the 2-norm of the residual, they monitor the process based on a ball $B^n(R)$ enclosing the residual set $D_r$. As a result, they do not generate alarms when there is an attack that results in residuals of the process at some time $t > 0$ which satisfy $r(t) \in B^n(R) \setminus D_r$. Thus, a limitation of the class of residual-based detection schemes monitoring the process with the 2-norm of the residual as the monitoring variable is their failure to distinguish between a residual $r(t) \in D_r$ and a residual $r'(t) \in B^n(R) \setminus D_r$ for some time $t$. As an alternative, another class of residual-based detection schemes that monitor the set-membership of the residuals may be considered of the form:*

$$h_s(t) = \begin{cases} 0 & r(t) \in D_r \\ 1 & r(t) \notin D_r \end{cases} \tag{2.15}$$

*for all time $t \geq 0$ with $h_s(t) = 1$ indicating anomalous process behavior. This class of detection schemes may be able to detect attacks that are undetectable with respect to the detection schemes considered in Eq. 2.11, i.e., attacks that lead to residuals such that $r'(t) \in B^n(R) \setminus D_r$. However, if there is an attack on the process that results in Eq. 2.14, then for all time $t \geq 0$, the residual of the process $r(t) \in D_r$, and the attack is undetectable. This makes Eq. 2.14 an important condition for undetectability of an attack.*

### 2.2.2 Minimum Invariant Set and Residual Bound Estimation

To verify the undetectability condition established in Eq. 2.13 and present a numerically implementable solution that can be leveraged for the controller screening, an approach to estimate $D_r$ and $D_{r_a}$ is presented in this section. The approach is adapted from a method to generate polytopic outer approximations of minimum invariant sets presented in the literature [79]. The computational method produces outer approximations of the sets $D_r$ and $D_{r_a}$. To address the numerical error associated with the approximations of the sets $D_r$ and $D_{r_a}$, a computationally verifiable condition for attack undetectability is derived that accounts for the maximum error associated with the approximations of the sets $D_r$ and $D_{r_a}$.

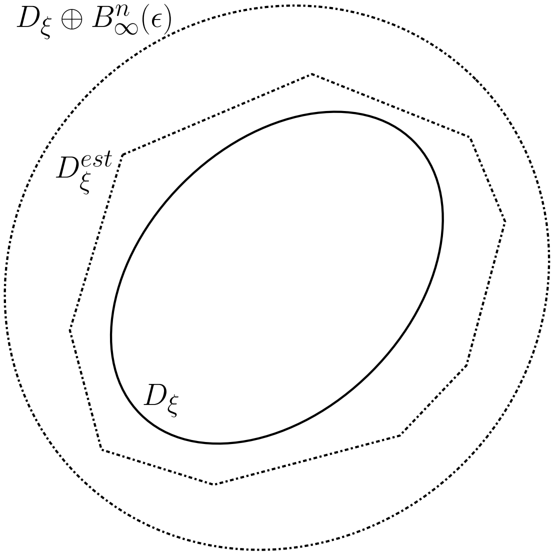


Fig. 2.2: The outer estimate ($D_\xi^{est}$) of the minimum invariant set ($D_\xi$) of the augmented closed-loop system.

To make the presentation self-contained, the method for generating a polytopic outer approximation [79] of the minimum invariant set is summarized here. Applying the method in Ref. 79 to the augmented closed-loop system (Eq. 2.6), the minimum invariant set represented by the infinite Minkowski sum $D_\xi = \bigoplus_{i=0}^{\infty} A_\xi^i B_\xi F$ is truncated to:

$$D_{\xi,s} \triangleq \bigoplus_{i=0}^{s-1} A_\xi^i B_\xi F$$

where $s$ is the truncation step for the infinite sum chosen such that $A_\xi^s B_\xi F \subseteq \alpha B_\xi F$, $\alpha \in [0,1)$ is a small positive number such that the estimate of the minimum invariant set, denoted by $D_\xi^{est}(\alpha, s)$, satisfies:

$$D_\xi^{est}(\alpha, s) \triangleq (1-\alpha)^{-1} D_{\xi,s}$$

where $D_\xi^{est}(\alpha, s)$ is convex, compact, and positive invariant set for the process (Eq. 2.6) with $0 \in D_\xi^{est}(\alpha, s)$ and $D_\xi \subseteq D_\xi^{est}(\alpha, s)$. Furthermore, the sum is truncated by evaluating $\alpha$ for a pre-determined estimation error $\epsilon$ so that $D_\xi \subseteq D_\xi^{est}(\alpha, s) \subseteq D_\xi \oplus B_\infty^n(\epsilon)$. Numerically, this involves incrementing $s$ starting from zero until $\alpha \leq \frac{\epsilon}{\epsilon + M(s)}$ where $M(s) \triangleq \min_\gamma \{\gamma \mid D_{\xi,s} \subseteq B_\infty^n(\gamma)\}$. In the remainder, the dependence of $\alpha$ and $s$ on $D_\xi^{est}$ is suppressed. Fig. 2.2 illustrates the bounds on polytopic estimate of the minimum invariant set $D_\xi^{est}$ which contains the minimum invariant set $D_\xi$, and is itself contained within the set $D_\xi \oplus B_\infty^n(\epsilon)$. The set $D_{\xi_a}$ may be analogously estimated for a given attack magnitude $\Lambda$ and the estimate is denoted by $D_{\xi_a}^{est}$. As $\epsilon \to 0$, the accuracy of invariant set estimates increases. However, the computational complexity of the estimation of $D_\xi^{est}$ increases as $\epsilon \to 0$ (c.f. Remark 1 [79]). Thus, the trade-off between the computational complexity and the accuracy of estimates needs to be managed when selecting the value of $\epsilon$.

The estimates of the residual sets, denoted by $D_r^{est}$ and $D_{r_a}^{est}$, can then be computed from the estimates of the minimum invariant sets:

$$\begin{aligned} D_r^{est} &= A_r D_\xi^{est} \oplus B_r F \\ D_{r_a}^{est} &= A_{r_a} D_{\xi_a}^{est} \oplus B_{r_a} F \end{aligned} \tag{2.16}$$

Furthermore, the bounds on the residual set estimates are:

$$\begin{aligned} D_r \subseteq D_r^{est} \subseteq D_r \oplus A_r B_\infty^n(\epsilon) \\ D_{r_a} \subseteq D_{r_a}^{est} \subseteq D_{r_a} \oplus A_{r_a} B_\infty^n(\epsilon) \end{aligned} \tag{2.17}$$

Thus, $D_r^{est}$ and $D_{r_a}^{est}$ are outer estimates of the sets $D_r$ and $D_{r_a}$ with error bounds described by radii of the transformed balls $A_r B_\infty^n(\epsilon)$ and $A_{r_a} B_\infty^n(\epsilon)$, respectively, similar to the illustration in Fig. 2.2. Furthermore, if $D_r \subseteq D_r^{est}$ and $D_{r_a} \subseteq D_{r_a}^{est}$, then:

$$R_a \leq R_a^{est} \tag{2.18a}$$

$$R \leq R^{est} \tag{2.18b}$$

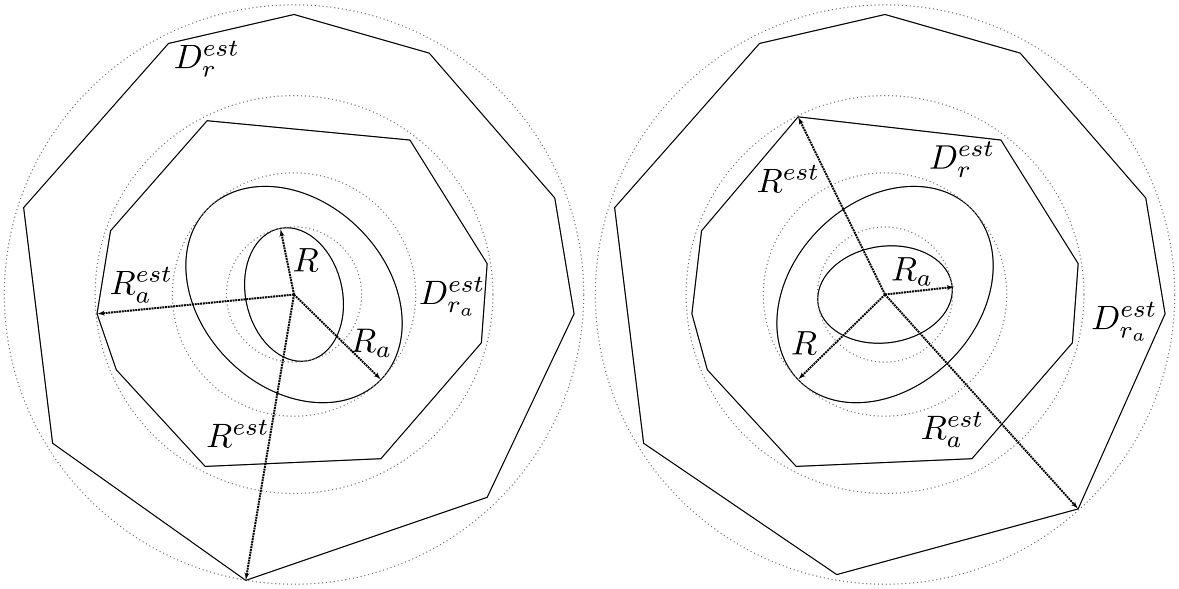
where $R_a^{est} := \max_{r' \in D_{r_a}^{est}} \|r'\|$ and $R^{est} := \max_{r' \in D_r^{est}} \|r'\|$. If there is an attack on the process such that the sets $D_r$ and $D_{r_a}^{est}$ satisfy $R_a^{est} \leq R$, then the attack is undetectable. However, since $D_r$ and $D_{r_a}$ cannot be computed in general, and $D_r^{est}$ and $D_{r_a}^{est}$ are outer approximations of $D_r$ and $D_{r_a}$, there is insufficient information to conclude that $R_a \leq R$ if $R_a^{est} \leq R^{est}$. Fig. 2.3 provides illustrative examples of cases where the result of testing for Eq. 2.13 using the residual set estimates ($D_r^{est}$ and $D_{r_a}^{est}$) may result in the test falsely indicating that the attack is undetectable with $R_a^{est} \leq R^{est}$ when it may actually be detectable with $R_a > R$ (e.g., Fig. 2.3a) or the test may indicate that the attack is detectable with $R_a^{est} > R^{est}$ when it actually may be undetectable with $R > R_a$ (e.g., Fig. 2.3b). This motivates the need to account for the estimation error to derive a numerically verifiable condition to determine if an attack is undetectable and is stated in the main result of the paper.

**Theorem 1.** *Consider the closed-loop process represented by the dynamics in Eq. 2.1 under a multiplicative sensor-controller link attack of magnitude $\Lambda \neq I$ with the controller in Eq. 2.4 using the state estimate from the observer in Eq. 2.3 and monitored by a detection scheme that fits the model for the class of residual-based detection scheme in Eq. 2.11. Let the closed-loop process be stable in the sense that all the eigenvalues of $A_\xi$ and $A_{\xi_a}$ (Eq. 2.6) are within the unit circle. If $D_r^{est}$ and $D_{r_a}^{est}$ are numerical estimates of residual sets computed based on Eq. 2.16 and $R_a^{est} \leq R_e^{est}$ where $R_e^{est} := \max_{r' \in D_{r_e}^{est}} \|r'\|$ and $D_{r_e}^{est} := D_r^{est} \ominus A_r B_\infty^n(\epsilon)$, then the attack is undetectable.*

From Theorem 1, the condition that must be verified to check for undetectability of an attack, given $K$ and $L$ is:

$$R_a^{est} \leq R_e^{est} \tag{2.19}$$

The subsequent corollary directly follows from similar arguments already established for potentially detectable attacks.



(a) (b)

Fig. 2.3: (a) An illustrative case showing false positive test result for undetectability with $R_a^{est} < R^{est}$, $R_a > R$. (b) An illustrative case showing false negative test result for undetectability with $R_a^{est} > R^{est}$, $R_a < R$.

**Corollary 1.** *Consider the closed-loop process represented by the dynamics in Eq. 2.1 under a multiplicative sensor-controller link attack of magnitude $\Lambda \neq I$ with the controller in Eq. 2.4 using the state estimate from the observer in Eq. 2.3 and monitored by a detection scheme that fits the model for the class of residual-based detection scheme in Eq. 2.11. Let the attack-free and the attacked closed-loop process be stable in the sense that all eigenvalues of $A_\xi$ and $A_{\xi_a}$ are within the unit circle. If $R_{a,e}^{est} := \max_{r' \in D_{r_{a,e}^{est}}} \|r'\| \geq R^{est}$ where $D_{r_{a,e}}^{est} := D_{r_a}^{est} \ominus (A_{r_a} B_\infty^n(\epsilon))$, then the attack is potentially detectable.*

Similarly, the following proposition is based on the arguments already established for detectable attacks.

**Proposition 2.** *Consider the closed-loop process represented by the dynamics in Eq. 2.1 under a multiplicative sensor-controller link attack of magnitude $\Lambda \neq I$ with the controller in Eq. 2.4 using the state estimate from the observer in Eq. 2.3 and monitored by a detection scheme that fits the model for the class of residual-based detection scheme in Eq. 2.11. Let the attack-free closed-loop process be stable in the sense that all eigenvalues* *of $A_\xi$ are within the unit circle. If the closed-loop process under attack is unstable, then the attack is detectable.*

Owing to the fact that $D_r$ and $D_{r_a}$ cannot be computed exactly, no conclusion can be made about the undetectability of an attack when the residual sets do not satisfy the detectability conditions in Eq. 2.19 and Corollary 1 with $R_a^{est} \geq R_e^{est}$ and $R_{a,e}^{est} \ngeq R^{est}$. Similarly, when the residual sets are such that they do not satisfy the detectability conditions with $R_{a,e}^{est} \leq R^{est}$ and $R_a^{est} \ngeq R_e^{est}$, no conclusion can be made about the potential detectability of the attack. These inconclusive test results occur when $|R_{a,e}^{est} - R^{est}| \leq \epsilon$ and $|R^{est} - R_e^{est}| \leq \epsilon'$ where $\epsilon > 0$ and $\epsilon' > 0$ are small numbers. From a practical perspective, the probability of a sequence of residuals taking values in the set $B^n(R_a) \setminus B^n(R)$ may be small since the size of the set $B^n(R_a) \setminus B^n(R)$ is small. Therefore, an inconclusive test result may mean that the attack is practically undetectable.

As discussed previously, Eq. 2.14 is an important condition for the undetectability of an attack, as it characterizes all attacks that result in the residual sequences under attack being indistinguishable from the attack-free residual sequences. From a practical perspective, it may be in the interest of a control designer, to specifically avoid the choices of controller and observer gains $K$ and $L$, that lead to satisfaction of Eq. 2.14. Numerically, this condition may be verified using the result from Section 1 of Ref. [80]. Specifically, if $D_{r_e}^{est}$ and $D_{r_a}^{est}$ are polytopes with $N$ and $M$ faces respectively, then they may be represented mathematically as:

$$D_{r_e}^{est} = \{r \mid a_{r_i}^T r \leq b_{ri},\ i = 1, 2, \ldots, N\}$$
$$D_{r_a}^{est} = \{r_a \mid {a_{r_a}}_j^T r_a \leq b_{r_a j},\ j = 1, 2, \ldots, M\}$$

The set $D_{r_a}^{est}$ is a subset of $D_{r_e}^{est}$ if and only if the following condition holds:

$$h_{D_{r_a}^{est}}(a_{r_i}) \leq b_{ri},\ \forall i \in \{1, 2, \ldots, N\} \tag{2.20}$$

where $h_{D_{r_a}^{est}}(a_{r_i})$ is the support function of $D_{r_a}^{est}$ evaluated at $a_{r_i}$ for a polytope. Eq. 2.20 can be evaluated numerically as the optimal value of the following linear program for each

$i \in \{1, 2, \ldots, N\}$:

$$h_{D_{ra}^{est}}(a_{r_i}) = \max_{r_a} a_{r_i}^T r_a$$

$$\text{s.t. } a_{r_a j}^T r_a \leq b_{r_a j}, j = 1, 2, \ldots M$$

Similar to the result presented in Theorem 1, if:

$$D{r_a}^{est} \subseteq D_{r_e}^{est} \tag{2.21}$$

the attack is undetectable. Additionally, if $D_r^{est} \subseteq D_{r_{a,e}}^{est}$, then the attack on the process will be potentially detectable.

**Remark 2.2.2.** *A special case of minimum invariant set estimation occurs when $A_\xi$ is a nilpotent matrix with index s, i.e., there exists an $s > 0$ such that $A_\xi^s = 0$. In this case, the minimum invariant set of the augmented closed-loop system can be computed exactly as $D_\xi = D_{\xi_s} \triangleq \bigoplus_{i=0}^{s-1} A_\xi^i B_\xi F$ (Theorem 3 [81]), and $D_\xi^{est} = D_\xi$ with $D_r^{est} = D_r$ leading to zero error in estimation of the residual sets. However, if the index of nilpotence s is a large number, to reduce the computational complexity, it may be preferable to estimate the set according to the algorithm presented previously for a non-nilpotent matrix $A_\xi$.*

### 2.2.3 Controller Screening Methodology

The controller screening methodology leverages the undetectability condition in Eq. 2.19. The algorithm presented herein considers the process represented by Eq. 2.1-2.2 with an observer of the form Eq. 2.3 and provides a tool for a control designer to check if the choice of controller and observer gains $K$ and $L$ render a multiplicative sensor-controller attack of given magnitude $\Lambda \neq I$ undetectable or potentially detectable. The controller gain $K$, the observer gain $L$, and the attack matrix $\Lambda$ are determined by the control designer prior to the screening. This predetermination may be based on stability and performance considerations of the process. Fig. 2.4 shows a flowchart of the screening algorithm.

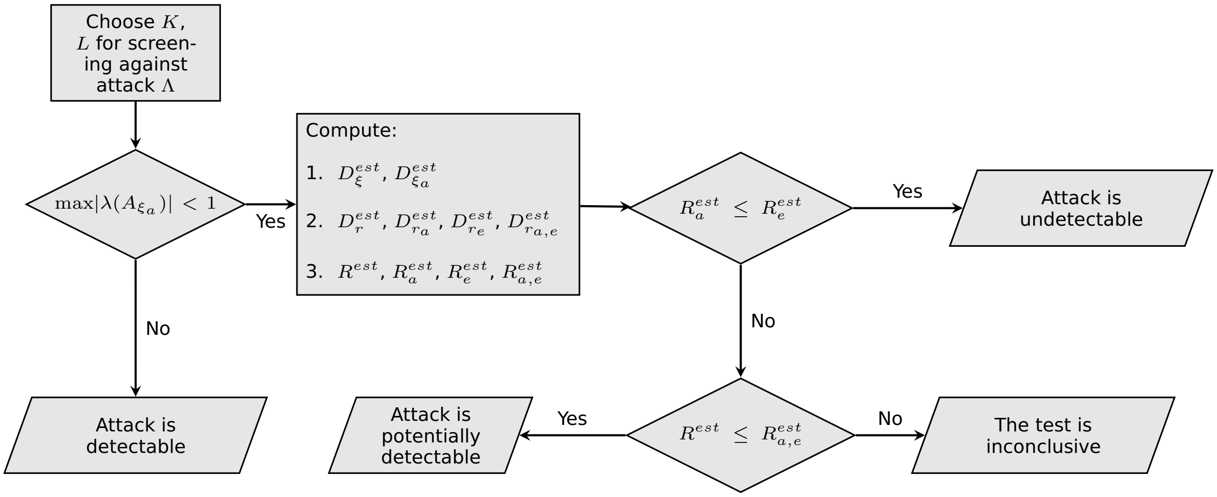

Choose $K$, $L$ for screening against attack $\Lambda$
$\max|\lambda(A_{\xi_a})| < 1$
Yes
No
Attack is detectable
Compute:
1. $D_{\xi}^{est}$, $D_{\xi_a}^{est}$
2. $D_r^{est}$, $D_{r_a}^{est}$, $D_{r_e}^{est}$, $D_{r_{a,e}}^{est}$
3. $R^{est}$, $R_a^{est}$, $R_e^{est}$, $R_{a,e}^{est}$
$R_a^{est} \leq R_e^{est}$
Yes
Attack is undetectable
No
$R^{est} \leq R_{a,e}^{est}$
Yes
Attack is potentially detectable
No
The test is inconclusive

Fig. 2.4: Flowchart for the controller design screening algorithm.

The algorithm is summarized by the following steps:

1. Check if $\max|\lambda(A_{\xi_a})| < 1$ for the augmented closed-loop system under attack in Eq. 2.6. If true, then go to Step 2. Else, the choice of $K$ and $L$ under a multiplicative attack with attack matrix $\Lambda$ will render the closed-loop process unstable. The attack is detectable. Terminate the screening algorithm.

2. Compute the estimates of the minimum invariant sets ($D_{\xi_a}{}^{est}$ and $D_{\xi}{}^{est}$) in the presence and absence of the attack for the augmented closed-loop system in Eq. 2.6.

3. Compute the outer estimates of the residual sets $D_r$ and $D_{r_a}$ using Eq. 2.17 as $D_r{}^{est}$ and $D_{r_a}{}^{est}$, respectively. From the outer estimates, compute the inner estimates of the residual sets $D_r$ and $D_{r_a}$ using Eq. A.1 and Corollary 1 as $D_{r_e}^{est}$, and $D_{r_{a,e}}^{est}$, respectively.

4. Compute the radii of the balls enclosing the sets $D_r^{est}$, $D_{r_a}{}^{est}$, $D_{r_e}^{est}$, and $D_{r_{a,e}}^{est}$ as $R^{est}$, $R_a^{est}$, $R_e^{est}$ and $R_{a,e}^{est}$, respectively.

5. Detectability verification:

   5.1 Check if Eq. 2.19 is satisfied, i.e., $R_a^{est} \leq R_e^{est}$. If true, the attack is undetectable. Else go to Step 5.2.

5.2 If $R^{est}_{a,e} \geq R^{est}$, then the attack is potentially detectable with the choice of $K$ and $L$. Else go to Step 5.3.

5.3 If $R^{est}_{a} > R^{est}_{e}$ or $R^{est}_{a,e} < R^{est}$, the test is inconclusive and the attack may be practically undetectable.

## 2.3 Application to Illustrative Processes

In this section, the proposed control parameter screening methodology is applied to two illustrative processes. All set computations within this section are performed using the MPT toolbox [82]. For the cases considered, the CUSUM detection scheme is used to verify the results of the controller screening methodology using closed-loop simulations. To track the total number of threshold breaches per simulation, the detection scheme is reset to 0 after each threshold breach, i.e., if $S(t) > \tau$, $S(t+1) = 0$.

### 2.3.1 Illustrative Process 1: Scalar Process

Consider an example process with the following dynamics:

$$x(t+1) = x(t) + u(t) + w(t)$$
$$y(t) = x(t) + v(t)$$

where $x(t)$, $y(t)$, and $u(t)$ are the state, output, and the manipulated input vectors, respectively. A controller of the form of Eq. 2.4 is used to regulate the process, and an observer of the form of Eq. 2.3 is used to estimate the process state.

The process disturbance $w(t) \in W$ and measurement noise $v(t) \in V$ satisfy $-1.5 \leq w(t) \leq 1.5$ and $-1.5 \leq v(t) \leq 1.5$, and are modeled as random variables drawn from a truncated distribution derived from a zero-mean Gaussian distribution with unit variance.

The subsequent sections present the screening of choices of $K$ and $L$ for an undetectable attack with $\Lambda = 0.9$. The minimum invariant set computations for the augmented state are performed with an error bound of $\epsilon = 5 \times 10^{-5}$.

#### 2.3.1.1 Undetectable Attack Case

In this case, a controller and observer gain of $K = 1.081$, and $L = 0.246$ are screened for potential to result in an undetectable attack with $\Lambda = 0.9$. The closed-loop augmented system is stable under an attack with $\max|\lambda(A_{\xi_a})| = 0.7847$. The minimum invariant set estimates, ${D_\xi}^{est}$ and ${D_{\xi_a}}^{est}$, along with an example closed-loop trajectory are shown in Fig. 2.5a and Fig. 2.5b, respectively. While $D_\xi^{est}$ and $D_{\xi_a}^{est}$ may be contained within one another, their relative positions and sizes may not be an indication of the detectability of an attack. It is their impact on the residual set sizes that governs the controller screening criterion. The estimate of the set of residuals under attack-free conditions is computed as:

$$ {D_r}^{est} = \{r' \mid \|r'\| \leq 9.0976\} \equiv \{r' \mid {A_{r_i}}^T r' \leq a_{r_i}, i = 1, 2\} $$

where $A_{r1} = -1$, $A_{r2} = 1$ and $a_{r1} = a_{r2} = 9.0976$. Similarly, the estimated set of residuals under attack can be computed as

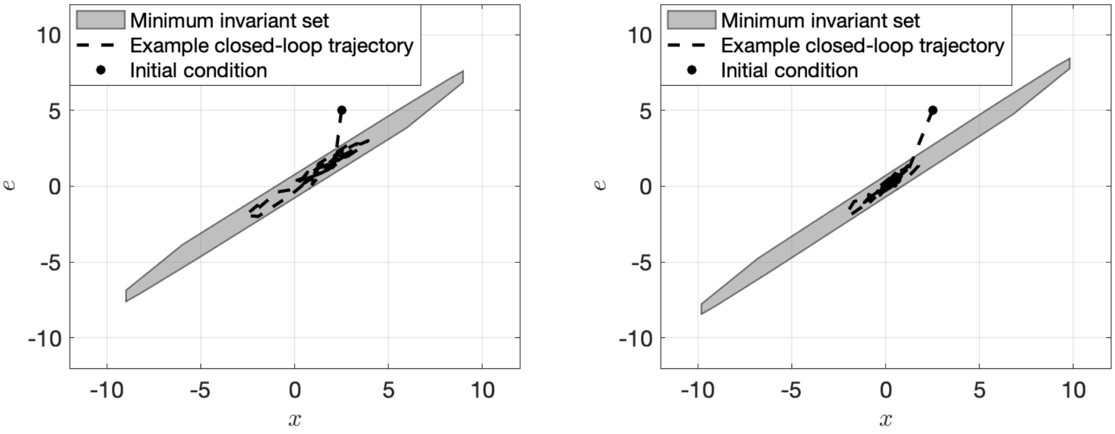
$$ {D_{r_a}}^{est} = \{r' \mid \|r'\| \leq 8.7976\} \equiv \{r' \mid {A_{r_{ai}}}^T r' \leq a_{r_{a_i}}, i = 1, 2\} $$

where $A_{r_{a_1}} = -1$, $A_{r_{a2}} = 1$ and $a_{r_{a1}} = a_{r_{a2}} = 8.7976$. The set $D_{r_e}^{est}$ is evaluated as:

$$ {D_{r_e}}^{est} = \{r' \mid \|r'\| \leq 9.0975\} \equiv \{r' \mid {A_{r_{ei}}}^T r' \leq a_{r_{e_i}}, i = 1, 2\} $$

where $A_{r_{e_1}} = -1$, $A_{r_{e2}} = 1$ and $a_{r_{e1}} = a_{r_{e2}} = 9.0975$. Based on this, $R_e^{est} = 9.075$ and $R_a^{est} = 8.7976$. Thus, $R_a^{est} < R_e^{est}$ with Eq. 2.19 satisfied, and the attack is undetectable based on the result of Theorem 1.

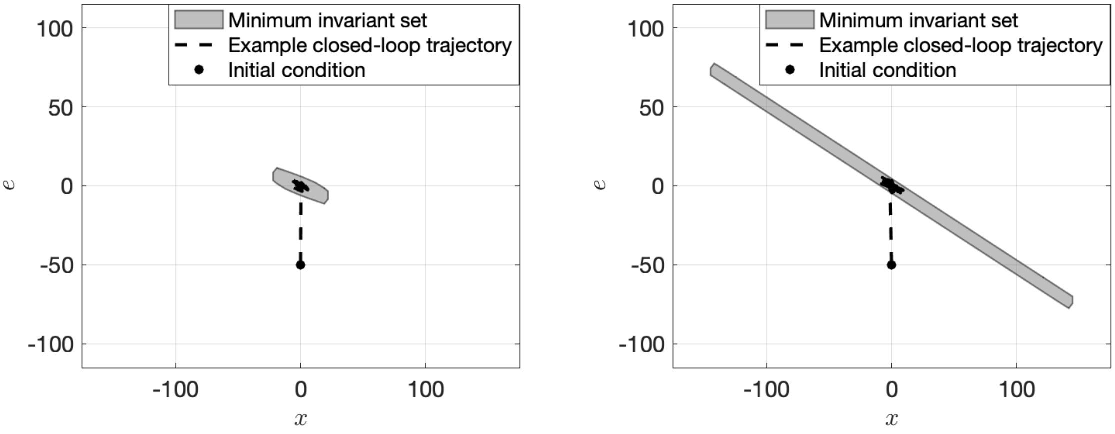
The undetectability of attack for this choice of controller and observer gains, and attack magnitudes is verified from closed-loop simulations of the process. The CUSUM parameter is selected as $b = R^{est} = 9.0976$. This ensures that the residual never breaches any arbitrarily small threshold in the absence of an attack. A small positive threshold of value $\tau = 0.1$ is used. One thousand simulations are performed of the attacked process, each of length $t = 5$ h for different realizations of the process disturbance and measurement noise. It is observed over these 1000 simulations that the CUSUM statistic never breaches the alarm threshold ($\tau = 0.1$), and remains at 0. As a result, the total number of threshold breaches over all simulations is zero, verifying that the attack is undetectable.



(a)

(b)

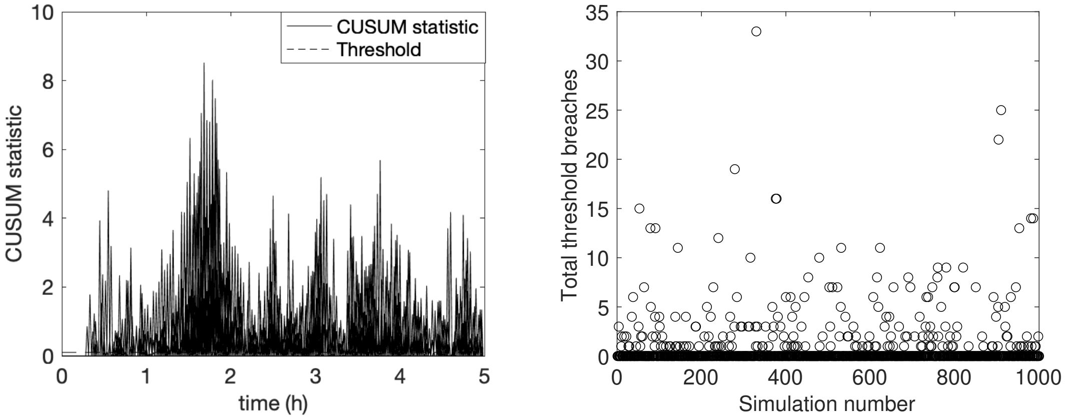
Fig. 2.5: The estimates of the minimum invariant set of (a) the attack-free augmented closed-loop system ($D_{\xi}{}^{est}$) and (b) the attacked augmented closed-loop system ($D_{\xi_a}^{est}$) for the scalar process in the undetectable attack case.



(a)

(b)

Fig. 2.6: The estimates of the minimum invariant set of (a) the attack-free augmented closed-loop system ($D_{\xi}{}^{est}$) and (b) the attacked augmented closed-loop system ($D_{\xi_a}^{est}$) for the scalar process in the potentially detectable attack case.



(a) (b)

Fig. 2.7: (a) The CUSUM statistic for 1000 simulations of the attacked scalar process for the potentially detectable attack case. (b) The number of threshold breaches per simulation for 1000 simulations of the attacked scalar process for the potentially detectable attack case.

#### 2.3.1.2 Potentially Detectable Attack Case

In this case a value of $K = 1.3$ and $L = 1.65$ are screened for the same attack magnitude ($\Lambda = 0.9$). The closed-loop augmented closed-loop system is stable under an attack with $\max|\lambda(A_{\xi_a})| = 0.9701$. The minimum invariant set estimates $D_\xi{}^{est}$ and $D_{\xi_a}{}^{est}$ are computed as shown in Fig. 2.6a and Fig. 2.6b. The estimated set of residuals under no attack is computed as:

$$D_r{}^{est} = \{r' \mid \|r'\| \leq 12.8572\} \equiv \{r' \mid A_{r_i}{}^T r' \leq a_{r_i}, i = 1, 2\}$$

where $A_{r1} = -1$, $A_{r2} = 1$ and $a_{r1} = a_{r2} = 12.8572$. The estimated set of residuals under attack is computed as

$$D_{r_a}{}^{est} = \{r' \mid \|r'\| \leq 92.9508\} \equiv \{r' \mid A_{r_{ai}}{}^T r' \leq a_{r_{a_i}}, i = 1, 2\}$$

where $A_{r_{a1}} = -1$, $A_{r_{a2}} = 1$ and $a_{r_{a1}} = a_{r_{a2}} = 92.9508$. Similarly, the set $D_{r_e}^{est}$ is computed as:

$$D_{r_e}{}^{est} = \{r' \mid \|r\| \leq 12.8571\} \equiv \{r' \mid A_{r_{ei}}{}^T r' \leq a_{r_{e_i}}, i = 1, 2\}$$

where $A_{r_{e1}} = -1$, $A_{r_{e2}} = 1$ and $a_{r_{e1}} = a_{r_{e2}} = 12.8571$. Based on this $R_a^{est} = 92.9508$ and $R_e^{est} = 12.8571$, indicating that $R_a^{est} \not\leq R_e^{est}$. Thus, Eq. 2.19 is not satisfied. To check for potential detectability per Corollary 1, the set $D_{r_{a,e}}^{est}$ is estimated as:

$$D_{r_{a,e}}{}^{est} = \{r' \mid \|r'\| \leq 92.9507\} \equiv \{r' \mid A_{r_{a,e_i}}{}^{T} r' \leq a_{r_{a,e_i}}, i = 1, 2\}$$

indicating that $R_{a,e}^{est} = 92.9507 > R^{est} = 12.8572$, meaning the attack is potentially detectable.

The potential detectability of attack for this choice of controller and observer gains, and attack magnitude is verified by closed-loop simulations of the process. The CUSUM parameter $b$ for this case is selected as $b = R^{est} = 12.8572$. This ensures that in the absence of an attack, the residual never breaches any arbitrarily small threshold and enables the choice of a small threshold of value $\tau = 0.1$. One thousand simulations are performed on the process, each of length $t = 5$ h for different realizations of noise.

As Fig. 2.7a and Fig. 2.7b demonstrate, the CUSUM statistic breaches the alarm threshold ($\tau = 0.1$) several times during some simulations over these 1000 simulations, resulting in the detection of the attack. However, there are some simulations during which the attack goes undetected with zero threshold breaches, demonstrating that the attack is not guaranteed to be detectable for all realizations of the process disturbance and measurement noise.

### 2.3.2 Illustrative Process 2: Continuous Stirred Tank Reactor

In this section, the application of the controller screening methodology is demonstrated using an illustrative process example consisting of a continuous stirred tank reactor (CSTR) where an exothermic, second-order reaction A $\longrightarrow$ B occurs. Under standard modeling assumptions, the material and energy balances describing the process dynamics are given by:

$$\begin{aligned} \frac{dC_A}{dt} &= \frac{F}{V}(C_{A0} + \Delta C_{A0} - C_A) - k_0 e^{\frac{-E}{RT}} C_A^2 \\ \frac{dT}{dt} &= \frac{F}{V}(T_0 + \Delta T_0 - T) - \frac{\Delta H k_0}{\rho_L C_p} e^{\frac{-E}{RT}} C_A^2 + \frac{Q}{\rho_L C_p V} \end{aligned} \tag{2.22}$$

where $C_A$ and $T$ are the concentration of the reactant and the temperature in the reactor, $F$ is the volumetric flow rate, $C_{A0}$ is the concentration of the reactant in the feed, $T_0$ is the temperature of the feed to the reactor, $\Delta H$ is the enthalpy of the reaction, $Q$ is the heat rate added or removed from the reactor, $\rho_L$ is the density of liquid in the reactor,

Table 2.1: Model parameters for the CSTR Process [1].

| | |
|---|---|
| Density | $\rho_L = 1000\,\mathrm{kg\,m^{-3}}$ |
| Heat capacity | $C_p = 0.231\,\mathrm{kJ\,kg^{-1}\,K^{-1}}$ |
| Flow rate | $F = 5.0\,\mathrm{m^3\,h^{-1}}$ |
| Reactor volume | $V = 1.0\,\mathrm{m^3}$ |
| Heat of reaction | $\Delta H = -1.15 \times 10^4\,\mathrm{kJ\,kmol^{-1}}$ |
| Activation energy | $E = 5.0 \times 10^4\,\mathrm{kJ\,kmol^{-1}}$ |
| Feed temperature | $T_0 = 300.0\,\mathrm{K}$ |
| Pre-expontential factor | $k_0 = 8.46 \times 10^6\,\mathrm{m^3\,kmol^{-1}\,h^{-1}}$ |
| Gas constant | $R = 8.314\,\mathrm{kJ\,kmol^{-1}\,K^{-1}}$ |
| Concentration of reactant $A$ in the feed | $C_{A0} = 4.0\,\mathrm{kmol\,m^{-3}}$ |

$V$ is the volume of the reactor, $C_p$ is the heat capacity, $E$ is the activation energy, and $k_0$ and $R$ are the pre-exponential factor and the gas constant, respectively. The model parameters are given in Table 2.1 and are taken from Ref. 1.

The control objective is to operate the CSTR around the open-loop stable steady-state where $C_{As} = 1.22\,\mathrm{kmol\,m^{-3}}$ and $T_s = 438.2\,\mathrm{K}$. The manipulated input of the process is the heat rate removed from or added to the reactor $Q$. The state variables of the process include the reactant concentration in the reactor and the reactor temperature. Full-state measurements are assumed to be available for feedback control. Deviation variables of the manipulated input and the states of the reactor are $x = [x_1\ x_2]^T = [C_A - C_{As}\ T - T_s]^T$ and $u = Q - Q_s$ where $Q_s = 0\,\mathrm{kJ\,h^{-1}}$ is the steady-state value of $Q$. The process is continuously perturbed due to additive disturbances bounded between $-10^{-3}$ $K$ and $10^{-3}$ $K$ in the feed temperature ($\Delta T_0$), and the additive disturbances in the feed concentration ($\Delta C_{A0}$) bounded between $-10^{-3}\,\mathrm{kmol\,m^{-3}}$ and $10^{-3}\,\mathrm{kmol\,m^{-3}}$.

The nonlinear process model in Eq. 2.22 is linearized around the desired operating steady-state. The linearized state-space model of the CSTR is given by:

$$\dot{x}(t) = \frac{dx}{dt} = A_c x(t) + B_c u(t) + G_c w(t)$$

where

$$A_c = \begin{bmatrix} -27.7051 & -0.4348 \\ 1130.3 & 16.64 \end{bmatrix} ; \; B_c = \begin{bmatrix} 0 \\ 0.0043 \end{bmatrix} ; \; G_c = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}$$

A zero-order hold discretization is applied on the linearized state-space model with a sampling period of $10^{-2}$ h resulting in the following discrete-time linear model similar to Eq. 2.1, with:

$$A = \begin{bmatrix} 0.7364 & -0.0041 \\ 10.6953 & 1.1560 \end{bmatrix} , \; B = \begin{bmatrix} -9.0708 \times 10^{-8} \\ 4.6741 \times 10^{-5} \end{bmatrix} , \; G = \begin{bmatrix} 0.0433 & -0.0001 \\ 0.2724 & 0.0540 \end{bmatrix}$$

The process is subject to additive measurement errors bounded between $-10^{-3}\,\mathrm{kmol\,m^{-3}}$ and $10^{-3}\,\mathrm{kmol\,m^{-3}}$ on the concentration sensor and additive measurement errors bounded between $-10^{-3}$ $K$ and $10^{-3}$ $K$ on the temperature sensor. A Luenberger observer of the form presented in Eq. 2.3 estimates the states of the CSTR with a feedback control law of the form of Eq. 2.4.

Expressing the system of equations in terms of the augmented state $\xi(t) = [x^T(t)\ e^T(t)]^T$ enables the computation of the minimum invariant set estimates and performing the controller screening for three different cases: (1) undetectable attack case: $K$ and $L$ that mask the detectability of an attack of magnitude 0.8 on the concentration sensor, (2) potentially detectable attack case-1: $K$ and $L$ that do not mask the detectability of an attack of magnitude 0.8 on the concentration sensor, and (3) potentially detectable attack case-2: $K$ and $L$ that do not mask the detectability of an attack of magnitude 0.9 on the temperature sensor. Furthermore, the minimum invariant sets for the process under attack and in the absence of the attack are computed using an error bound of $\epsilon = 5 \times 10^{-5}$.

While the controller screening methodology is applied to the linearized discrete-time model of the process, the closed-loop simulations used to verify the result from the screening methodology are performed using the nonlinear model of the CSTR with all process disturbances and measurement noise modeled as random variables drawn from a truncated distribution derived from a zero-mean Gaussian distribution with unit variance. The simulations emulate the conditions in a real-time PCS where the process itself is nonlinear, however, the control law is linear and is implemented in a digital computer with a zero-order hold. To integrate the ordinary differential equations describing the nonlinear closed-loop system, an explicit Euler's method with a step size of $10^{-4}$ h is used.

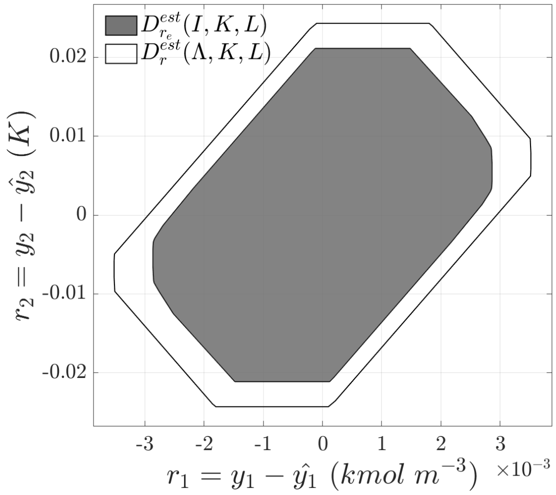


Fig. 2.8: The outer estimate of the residual set $(D^{est}_{r_a})$ of the attacked CSTR and the inner estimate of the residual set $(D^{est}_{r_e})$ of the attack-free CSTR in the undetectable attack case.

#### 2.3.2.1 Undetectable Attack Case

For this case, the controller choice screened for its potential to mask a multiplicative attack of magnitude represented by $\Lambda = \text{diag}(0.8, 1)$ is one with the controller poles at [0.2 0.3] and observer poles at $[-0.5\ 0.5]$. The eigenvalue of the attacked augmented closed-loop system with the largest magnitude, $\max|\lambda(A_{\xi_a})| = 0.5109$ indicating that closed-loop stability of the process is preserved under attack. The values of $R^{est}_e$ and $R^{est}_a$ are evaluated as $R^{est}_a = 0.0244 < R^{est}_e = 0.0212$ indicating that the attack is undetectable. Fig. 2.8 illustrates the outer estimate of the residual set for the attacked process and the inner estimate of the residual set for the attack-free process.

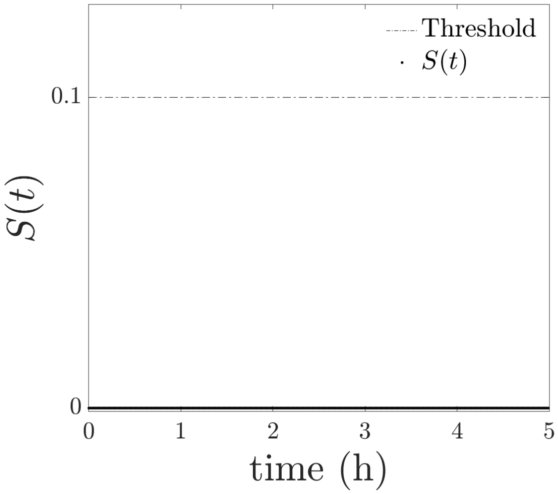


Fig. 2.9: CUSUM statistic values observed over 1000 simulations of the attacked CSTR process for the undetectable attack case.

The undetectability of the attack is verified by running one thousand simulations of the nonlinear sampled-data process model for many realizations of the process disturbance and measurement noise. The CUSUM detection scheme was tuned with the choice of $b = R^{est} = 0.0244$ and $\tau = 0.1$. As illustrated in Fig. 2.9, the CUSUM statistic over these one thousand simulations remains at 0 and the threshold is never breached. This confirms that the attack is undetectable.

#### 2.3.2.2 Potentially Detectable Attack Case

For this case, an attack of magnitude 0.9 on the temperature sensor is considered with an attack matrix $\Lambda = diag(1, 0.9)$. The controller and observer gains for the process are chosen with the controller poles at $[0.3 \ -0.1]$ and observer poles at $[0.2 \ 0.2]$. The eigenvalue of the attacked augmented closed-loop system with the maximum magnitude, $\max|\lambda(A_{\xi_a})| = 0.5365$ indicating that the closed-loop stability of the process is preserved under attack. The values of $R_e^{est}$ and $R_a^{est}$ are evaluated as $R_a^{est} = 0.0277 > R_e^{est} = 0.0161$. Similarly, based on the sets $D_{r_{a,e}}^{est}$ and $D_r^{est}$ (Fig. 2.10), $R_{a,e}^{est}$ and $R^{est}$ are evaluated as $R_{a,e}^{est} = 0.0277 > R^{est} = 0.0161$ indicating that the attack is potentially detectable.

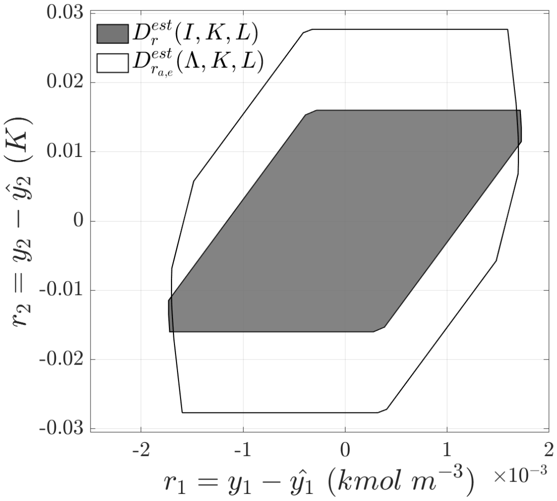


Fig. 2.10: The outer estimate of the residual set ($D_r^{est}$) of the attack-free CSTR and the inner estimate of the residual set ($D_{r_{a,e}}^{est}$) of the attacked CSTR in the potentially detectable attack case.

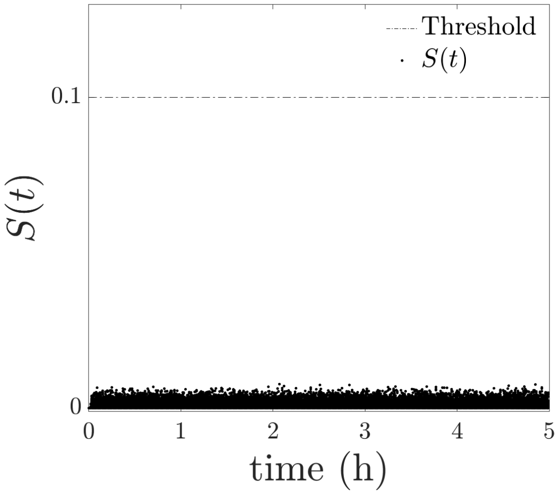


Fig. 2.11: CUSUM statistic values observed over 1000 simulations of the attacked CSTR process for the potentially detectable attack case.

To verify the results on a closed-loop simulation of the attacked process, one thousand simulations of the attacked nonlinear process model are performed for many realizations of the process disturbance and measurement noise. The CUSUM detection scheme was tuned with the choice of parameters $b = R^{est} = 0.0161$ and $\tau = 0.1$. Fig. 2.11 shows that even with a choice of control parameters that do not mask the detectability of the attack, the threshold is never breached over these one thousand simulations. However, unlike the previous case, the CUSUM statistic does not remain at zero (Fig. 2.9) and has a mean of 0.0064 and a variance of $1.8464 \times 10^{-5}$ over one thousand simulations. With a lower threshold (e.g., $\tau = 10^{-5}$), the attack is detectable.

## 2.4 Conclusions

In this chapter, an approach to incorporating the detectability of a cyberattack into the existing controller design criteria was presented. First, attacks were classified as undetectable and potentially detectable based on their impact on a class of residual-based attack detection schemes tuned to generate zero alarms during attack-free operation. Then, for a given magnitude of multiplicative sensor-controller attack, controller and observer gains, a residual set-based condition for undetectability was obtained. Subsequently, leveraging the characterized undetectability condition, a controller screening methodology to aid the control designer in identifying controller and observer gains that mask the detectability of an attack of a certain magnitude was presented. Finally, the application of the proposed controller screening methodology was demonstrated using two examples, including a chemical process example with nonlinear dynamics. The chemical process example demonstrated the potential applicability of the proposed methodology to nonlinear processes.

# Chapter 3

# Active Multiplicative Cyberattack Detection Utilizing Controller Switching for Process Systems

In this chapter, the relationship between closed-loop stability, control system parameters, and attack detectability for a residual-based detection scheme is rigorously characterized. The characterization is used to identify a set of control system parameters (called "attack-sensitive" parameters) under which a multiplicative sensor-controller link attack can destabilize the closed-loop system. The attack-sensitive control system parameters are selected such that they can enhance the ability to detect attacks, but can also degrade the performance of the attack-free closed-loop system. A novel active attack detection methodology employing control system parameter switching is developed to balance the tradeoff between attack detection and closed-loop performance. The controller switches between the nominal control system parameters, chosen based on standard control design criteria, and the attack-sensitive parameters with the proposed detection method. The application of the active detection method is demonstrated using simulations of a chemical process example.

## 3.1 Preliminaries

### 3.1.1 Notation and Definition

For a vector $x \in \mathbb{R}^n$, its Euclidean norm is denoted by $\|x\|$, and its infinity norm is denoted by $\|x\|_\infty$. The closed Euclidean ball and infinity ball centered at the origin with radius $R > 0$ are denoted by $B^n(R) := \{x \in \mathbb{R}^n \mid \|x\| \leq R\}$ and $B^n_\infty := \{x \in \mathbb{R}^n \mid \|x\|_\infty \leq R\}$. For a compact set $D \subset \mathbb{R}^n$, $R_D$ denotes the minimal radius of the Euclidean ball enclosing the set, i.e., $R_D := \max\limits_{x \in D} \|x\|$. For a set $D \subset \mathbb{R}^n$, the linear transformation of the set is denoted by $AD := \{Ax \mid x \in D\}$. Given two nonempty sets $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^n$, their Minkowski sum is defined as $X \oplus Y = \{x + y \mid x \in X, y \in Y\}$. For matrices, diag$(\beta_1, \beta_2, \ldots, \beta_n)$ represents an $n \times n$ diagonal matrix with diagonal elements $\beta_1, \beta_2, \ldots, \beta_n$, $I$ represents the identity matrix of appropriate dimensions, and $\lambda_i(A)$ is the $i^{th}$ eigenvalue of the matrix $A$. Sequences are denoted with boldface letters, i.e., $\mathbf{d} := \{d(0), d(1), d(2), \ldots\}$ where $d(t) \in \mathbb{R}^n$ for all $t \geq 0$. For the discrete-time linear system: $z(t+1) = Az(t) + v(t)$, where $z(t) \in \mathbb{R}^n$, $v(t) \in V$ for all $t \geq 0$, and $V$ is a compact set, a set $D_z \subset \mathbb{R}^n$ is said to be robust positively invariant if $z(t) \in D_z$ implies that $z(t+1) \in D_z$ for any $v(t) \in V$. A set $M_z \subset \mathbb{R}^n$ is said to be a minimum robust positively invariant set if $M_z$ is contained within every closed robust positively invariant set [77]. For simplicity of presentation, the minimum robust positively invariant set will be referred to as the minimum invariant set in this chapter.

### 3.1.2 Class of Processes and Control System Design

Processes modeled by discrete-time linear time-invariant systems and subject to bounded process disturbances and bounded measurement noise are considered:

$$x(t+1) = Ax(t) + Bu(t) + Gw(t) \tag{3.1}$$

where $x(t) \in \mathbb{R}^{n_x}$ is the process state vector, $u(t) \in \mathbb{R}^{n_u}$ is the manipulated input vector, $w(t) \in W \subset \mathbb{R}^{n_w}$ is the bounded process disturbance vector, and the set $W$ is assumed to be a (compact) polytope containing the origin. Without loss of generality, the initial time is taken to be zero. The matrices $A$, $B$, and $G$ are of appropriate dimensions. The value of the measured output received by the controller may be corrupted by a multiplicative sensor-controller link attack. The measured output is modeled by:

$$y(t) = \Lambda(Cx(t) + v(t)) \tag{3.2}$$

where $y(t) \in \mathbb{R}^{n_y}$ is the potentially falsified output vector received by the controller, $v(t) \in V \subset \mathbb{R}^{n_y}$ is the measurement noise vector, the set $V$ is assumed to be a (compact) polytope containing the origin, and $\Lambda$ is the matrix modeling multiplicative sensor-controller link attack on the process. The matrix $C$ is of appropriate dimensions. The matrix $\Lambda$ is referred to as the attack magnitude where $\Lambda \neq I$ indicates the presence of an attack on the process and $\Lambda = I$ indicates the absence of an attack.

The matrix pair $(A, B)$ is assumed to be controllable, and the matrix pair $(A, C)$ is assumed to be observable. A Luenberger observer is used to estimate the process states and is given by:

$$\begin{aligned} \hat{x}(t+1) &= A\hat{x}(t) + Bu(t) + L(y(t) - \hat{y}(t)) \\ \hat{y}(t) &= C\hat{x}(t) \end{aligned} \tag{3.3}$$

where $\hat{x}(t) \in \mathbb{R}^{n_x}$ is the estimated state vector, $\hat{y}(t) \in \mathbb{R}^{n_y}$ is the estimated output vector, and $L \in \mathbb{R}^{n_x \times n_y}$ is the observer gain selected so the eigenvalues of $A - LC$ are within the unit circle. Without loss of generality, the desired operating steady-state for the process is assumed to be the origin. To steer the process state to the origin, a linear feedback control law is used:

$$u(t) = -K\hat{x}(t) \tag{3.4}$$

where $K \in \mathbb{R}^{n_u \times n_x}$ is the controller gain, selected such that the eigenvalues of $A - BK$ are within the unit circle.

The estimation error, defined as $e(t) = x(t) - \hat{x}(t)$, and the estimation error dynamics are given by:

$$e(t+1) = L(I - \Lambda)Cx(t) + (A - LC)e(t) + Gw(t) - L\Lambda v(t) \tag{3.5}$$

To analyze the stability of the overall closed-loop process consisting of the process in Eqs. 3.1-3.2 with the feedback control law in Eq. 3.4 using the estimated state from the observer in Eq. 3.3, an augmented state vector $\xi(t) = [x^T(t)\ e^T(t)]^T$ is defined. The augmented state dynamics are given by:

$$\xi(t+1) = \underbrace{\begin{bmatrix} (A-BK) & BK \\ L(I-\Lambda)C & (A-LC) \end{bmatrix}}_{:=A_\xi(\Lambda,K,L)} \xi(t) + \underbrace{\begin{bmatrix} G & 0_{n_x \times n_y} \\ G & -L\Lambda \end{bmatrix}}_{:=B_\xi(\Lambda,K,L)} d(t) \tag{3.6}$$

where $d(t) := \begin{bmatrix} w^T(t) & v^T(t) \end{bmatrix}^T \in F$ is the augmented disturbance and measurement noise vector, and $F := \left\{ \begin{bmatrix} w \\ v \end{bmatrix} \mid w \in W, v \in V \right\}$ is the set of disturbances. Here, $A_\xi(\Lambda, K, L)$ and $B_\xi(\Lambda, L)$ are the system matrices for the augmented state dynamics. In the remainder, the admissible set of disturbance and measurement noise sequences is denoted by $\mathcal{F} := \{\mathbf{d} \mid d(t) \in F,\ \forall\ t \geq 0\}$.

Given that chemical processes are typically operated at steady-state for long periods, all analyses in the present section focus on the process operating at its steady-state, i.e., after the augmented state of the closed-loop process has converged to its terminal set, which is the minimum invariant set. The minimum invariant set for the augmented system in Eq. 3.6 when $\max_i |\lambda_i(A_\xi(\Lambda, K, L))| < 1$ may be expressed as the infinite Minkowski sum [77]:

$$D_\xi(\Lambda, K, L) = \bigoplus_{i=0}^{\infty} A_\xi^i(\Lambda, K, L) B_\xi(\Lambda, L)\ F \tag{3.7}$$

Based on Eq. 3.7, the minimum invariant set of the augmented closed-loop system is dependent on the attack matrix $\Lambda$, the controller gain $K$, and the observer gain $L$. For simplicity, the process operated at steady-state refers to the system of Eq. 3.6 after the augmented state has converged to the minimum invariant set, i.e., $\xi(t) \in D_\xi(\Lambda, K, L)$ implying that $\xi(t+1) \in D_\xi(\Lambda, K, L)$. For the remainder, the term closed-loop process refers to the process described by Eqs. 3.1-3.2 under the feedback law given by Eq. 3.4 using the estimates of states generated by the observer in Eq. 3.3.

## 3.2 Active Multiplicative Attack Detection Utilizing Controller Switching

In this section, the residual-based detection scheme considered is discussed, and an approach for detectability-based classification of attacks is presented. Then, theoretical results characterizing the relationship between closed-loop stability, control system parameter selection, and the detectability of an attack with respect to the residual-based detection scheme considered, are presented. Leveraging the characterization, an active attack detection methodology using occasional control system parameter switching is developed.

### 3.2.1 Residual-Based Detection Scheme and Attack Detectability

For the closed-loop process, the residual vector ($r(t)$) is defined as the difference between the output ($y(t)$) and its estimate generated by the observer ($\hat{y}(t)$), i.e.,

$$r(t) := y(t) - \hat{y}(t)$$

Writing the residual in terms of the augmented state ($\xi(t)$) and the disturbance vector ($d(t)$) yields:

$$r(t) = \underbrace{\begin{bmatrix}(\Lambda - I)C & C\end{bmatrix}}_{=:A_r(\Lambda)} \xi(t) + \underbrace{\begin{bmatrix}0_{n_y \times n_w} & \Lambda\end{bmatrix}}_{=:B_r(\Lambda)} d(t) \tag{3.8}$$

When $A_\xi(\Lambda, K, L)$ has eigenvalues that lie within the unit circle and $F$ is compact, $D_\xi(\Lambda, K, L)$ is forward invariant [77] and compact (Sec. 4 in [80]), and the residual is ultimately bounded within a terminal set. From Eq. 3.8, the residual terminal set is given by:

$$D_r(\Lambda, K, L) = A_r(\Lambda) D_\xi(\Lambda, K, L) \oplus B_r(\Lambda) F \tag{3.9}$$

For every $\xi(t) \in D_\xi(\Lambda, K, L)$ and $\mathbf{d} \in \mathcal{F}$, all possible realizations of the residual will be contained within its terminal residual set, i.e., $r(t) \in D_r(\Lambda, K, L)$. Based on Eq. 3.8, in the absence of an attack ($\Lambda = I$), the residual is dependent on the estimation error (Eq. 3.5) and the disturbance ($d(t)$). However, in the presence of a multiplicative sensor-controller link attack ($\Lambda \neq I$), the residual is also coupled to the process state. In addition to its dependence on the disturbance set $F$, the minimum invariant set is dependent on both the controller gain ($K$) and the observer gain ($L$). This is true for both the attack-free and the attacked process. However, the dependency of the terminal residual set on the controller and observer gains varies for the attack-free and the attacked processes. Specifically, the attack-free terminal residual set is dependent on the the observer gain only, whereas the attacked terminal residual set is dependent on the both the controller gain and the observer gain. Nonetheless, to maintain uniformity of notation, $D_r(I, K, L)$ is used to represent the attack-free terminal residual set even though the terminal residual set is independent of $K$ when $\Lambda = I$.

Residual-based anomaly detection schemes are model-based detection schemes that are commonly used for process monitoring [83–87]. These detection schemes monitor the process without using external intervention. Consequently, they are passive detection schemes. Two types of residual-based detection schemes commonly employed for cyber-attack detection are the $\chi^2$ and CUSUM detection schemes [21, 38]. Both schemes are scalar detection schemes in the sense that their output values are scalar values. To monitor changes in the residual behavior over time, the schemes may be formulated using the 2-norm of the residual vector as the input driving the detector output (see, for example, [22] for further discussion on this point). To tune the detector to raise zero false alarms when the process is operating at steady-state, the tuning must account for the fact that the maximum achievable value of the 2-norm of the residual is equal to the radius of the ball enclosing the residual terminal set, i.e., $\|r(t)\| \leq R_{D_r}(I, K, L)$ where is $R_{D_r}(I, K, L)$ is the minimum radius of the 2-norm ball enclosing the residual terminal set ($D_r(I, K, L) \subseteq B^{n_y}(R_{D_r}(I, K, L))$) [22]. A limitation of such detection schemes is that they do not account for the shape of the terminal residual set of the attack-free closed-loop process $D_r(I, K, L)$. For example, if the residual of the attacked closed-loop process is such that it is outside the terminal residual set of the attack-free closed-loop process but bounded within the 2-norm ball enclosing the terminal residual set of the attack-free closed-loop process ($r(t) \in B^{n_y}(R_{D_r}(I, K, L)) \setminus D_r(I, K, L)$), the 2-norm residual-based detection schemes will not detect the attack. To overcome this limitation, a set membership-based detection scheme is considered. Specifically, the detection scheme considered is given by:

$$z(t) = \begin{cases} 0, & r(t) \in D_r(I, K, L) \\ 1, & r(t) \notin D_r(I, K, L) \end{cases} \tag{3.10}$$

where $z(t)$ represents the output of the detection scheme. An output of $z(t) = 0$ indicates normal process operation (no attack detection), and $z(t) = 1$ indicates that there an attack is detected. Since the set membership-based detection scheme does not use external intervention to monitor the process, it is a passive detection scheme.

Cyberattacks may be classified based on the ability of the detection scheme in Eq. 3.10 to detect the attack. For the closed-loop process operated at steady-state monitored by the detection scheme in Eq. 3.10, an attack is said to be detected at time $t_d$ if $r(t_d) \notin D_r(I, K, L)$ with the output of the detection scheme $z(t) = 1$. An attack is defined as a detectable attack with respect to the detection scheme in Eq. 3.10 if the attack is detected in finite time for all $\xi(0) \in D_\xi(\Lambda, K, L)$ and $\mathbf{d} \in \mathcal{F}$. If the attack renders the closed-loop process unstable, then by convention, the set $D_\xi(\Lambda, K, L)$ is taken to be the Euclidean space $\mathbb{R}^{2n_x}$. An attack is defined as an undetectable attack with respect to the detection scheme in Eq. 3.10, if the residual of the attacked closed-loop process satisfies $r(t) \in D_r(I, K, L)$ for all $t \geq 0$ for all $\xi(0) \in D_\xi(\Lambda, K, L)$ and $\mathbf{d} \in \mathcal{F}$. Finally, an attack is defined as potentially detectable with respect to the detection scheme in Eq. 3.10, if the attack is neither detectable nor undetectable. The set of initial conditions considered is $D_\xi(\Lambda, K, L)$ because steady-state operation is considered. For some initial conditions in $D_\xi(\Lambda, K, L)$, the attack is detected immediately by the detection scheme in Eq. 3.10. However, this does not imply that the attack is detectable, as the attack needs to be detected in finite-time for all initial conditions in $D_\xi(\Lambda, K, L)$. While the definitions for attack detectability with respect to the detection scheme in Eq. 3.10 are valid for any attack, multiplicative sensor-controller link attacks are considered in the present section. Owing to the process disturbances and measurement noise, the augmented process states of the stable closed-loop process (Eq. 3.6) are ultimately bounded within its minimum invariant set. Thus, the notion of closed-loop stability considered is ultimate boundedness of the augmented state of the closed-loop process. The closed-loop process in Eq. 3.6 is considered to be unstable if $\|\xi(t)\| \to \infty$ as $t \to \infty$.

To motivate the proposed active detection methodology, the relationship between closed-loop stability and detectability is analyzed first. Proposition 3 establishes a relationship between the undetectability of a multiplicative attack and the terminal residual sets of the attack-free and attacked closed-loop process.

**Proposition 3.** *Consider the closed-loop process operated at steady-state with control system parameters $(K, L)$ under a multiplicative sensor-controller link attack of magnitude $\Lambda$. If the attack is such that the closed-loop process remains stable, i.e., the eigenvalues of $A_\xi(\Lambda, K, L)$ lie within the unit circle, the multiplicative attack is undetectable with respect to the detection scheme in Eq. 3.10, if and only if $D_r(\Lambda, K, L) \subseteq D_r(I, K, L)$.*

From Proposition 3, the question may arise on whether $D_r(\Lambda, K, L) \nsubseteq D_r(I, K, L)$ and/or a conventional condition for instability, i.e., $\max_i |\lambda_i(A_\xi(\Lambda, K, L)| > 1$, are sufficient conditions for a detectable attack. However, these conditions alone are not sufficient conditions for a detectable attack, and can only be used to guarantee potential detectability of an attack, which is stated in the next proposition.

**Proposition 4.** *Consider the closed-loop process operated at steady-state with control system parameters $(K, L)$ under a multiplicative sensor-controller link attack of magnitude $\Lambda$. If the attack is such that (1) the attacked closed-loop process is stable with the eigenvalues of $A_\xi(\Lambda, K, L)$ within the unit circle, and $D_r(\Lambda, K, L) \nsubseteq D_r(I, K, L)$, or (2) the attacked closed-loop process is such that $\max_i |\lambda_i(A_\xi(\Lambda, K, L)| > 1$, then the attack is potentially detectable with respect to the detection scheme in Eq. 3.10.*

If the closed-loop process under an attack is unstable such that $\|\xi(t)\| \to \infty$ as $t \to \infty$ the attack will be detected in finite time, if an additional observability condition is satisfied. This result is formally stated in Proposition 5.

**Proposition 5.** *Consider the closed-loop process with control system parameters $(K, L)$ under a multiplicative attack of magnitude $\Lambda \neq I$. Let the control system parameters $(K, L)$ stabilize the attack-free closed-loop process. If the attack renders the closed-loop* *process unstable in the sense that $\|\xi(t)\| \to \infty$ as $t \to \infty$ and the pair $(A_\xi(\Lambda, K, L), A_r(\Lambda))$ is observable, the attack is detected in finite time with respect to the detection scheme in Eq. 3.10.*

The only assumption made about the process disturbance and measurement noise is that they are bounded. Even if some eigenvalues of $A_\xi(\Lambda, K^*, L^*)$ or $A_\xi(\Lambda, K_\Lambda, L_\Lambda)$ are outside the unit circle, this assumption does not exclude potential realizations of the disturbance and measurement noise that results in the augmented state remaining bounded for all times. These are cases where the disturbance can effectively act as a stabilizing input in the sense that the state remains bounded for all time, i.e., $\limsup_{t\to\infty} \|\xi(t)\|$ is finite. In practice, disturbances are exogenous inputs and are not expected to stabilize a process. The condition $\max_i |\lambda_i(A_\xi(\Lambda, K, L))| > 1$ is a necessary, but not sufficient, condition for the type of closed-loop instability considered in this section. Closed-loop instability cannot be verified solely by checking the eigenvalues of $A_\xi(\Lambda, K, L)$. Nevertheless, a multiplicative attack is said to be destabilizing if the eigenvalues of $A_\xi(\Lambda, K, L)$ are outside the unit circle ($\max_i |\lambda_i(A_\xi(\Lambda, K, L))| > 1$) to highlight that the attack is responsible for destabilization.

### 3.2.2 Active Attack Detection Methodology

Traditional control system design approaches use closed-loop stability, performance, and robustness to uncertainty as criteria to determine the control system design [73–75]. Although attack detectability is linked to the control system design (Section 3.2.1), traditional design methods do not consider cyberattack detectability and may result in selecting control system parameters that mask the cyberattack in the sense that a cyberattack goes undetected with these parameters. From an attack detection standpoint (Proposition 5), selecting control system parameters that are "sensitive" to cyberattacks, in the sense that the closed-loop process is rendered unstable by the attack, may be preferred. However, sustained operation with these control system parameters may not be desirable because the closed-loop performance may be worse than that achieved under parameters determined by traditional design approaches. To manage the trade-off between attack detection and closed-loop performance, the proposed active detection methodology utilizes occasional switching from the nominal control system parameters, determined by traditional design approaches, to the so-called attack-sensitive parameters. Control system parameter switching is one form of active detection that may be considered, owing to the link between control system parameters and attack detectability established in Section 3.2.1.

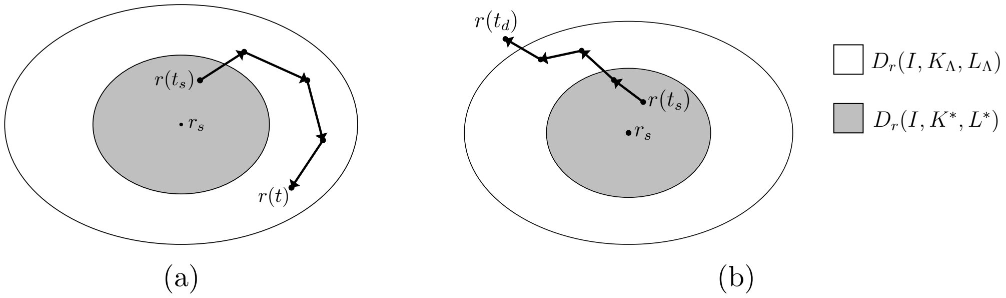


Fig. 3.1: (a) An example residual trajectory for the attack-free closed-loop process with the control system switch from nominal parameters to attack-sensitive parameters occurring at $t_s$. (b) An example residual trajectory for the attacked closed-loop process with the control parameter switch occurring at $t_s$ where the attack is detected at $t_d$.

The nominal parameters are denoted by $(K^*, L^*)$, while the attack-sensitive parameters are denoted by $(K_\Lambda, L_\Lambda)$. With the active detection methodology, the control system parameters switch from the nominal parameters to the attack-sensitive parameters at $t_s$. After the control system switches from the nominal parameters to attack-sensitive parameters occurs, the process is operated over a period $T_c > 0$ with the attack-sensitive parameters. Under attack-free operations (Fig. 3.1a), the residual trajectory after the switch will evolve in the terminal residual set of the attack-free closed-loop process with attack-sensitive parameters $D_r(I, K_\Lambda, L_\Lambda)$. After the period $T_c$ elapses, the control system switches back to the nominal parameters. In the presence of a multiplicative attack, the residual trajectory may evolve outside the terminal residual set of the attack-free closed-loop process with attack-sensitive parameters (Fig. 3.1b) resulting in the attack being detected.

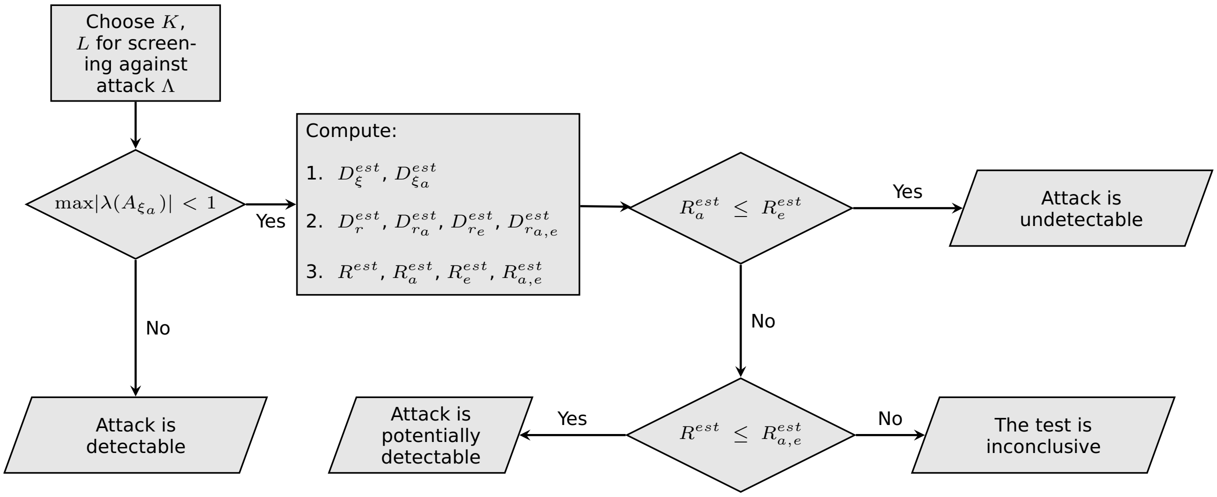
Under the active detection methodology, the control system parameters vary over time. The detection scheme needs to account for this change because the residual terminal set under attack-free operation depends on the controller and observer gains. Therefore, the detection scheme is modified as follows:

$$z(t) = \begin{cases} 0, & r(t) \in D_r(I, K(t), L(t)) \\ 1, & r(t) \notin D_r(I, K(t), L(t)) \end{cases} \tag{3.11}$$

where $K(t)$ is the controller gain used at time step $t$, and $L(t)$ is the observer gain at time step $t$, $z(t) = 0$ indicates anomaly-free operation, and $z(t) = 1$ indicates anomalous process operation. For the closed-loop process with the nominal parameters, $(K(t), L(t)) = (K^*, L^*)$, and for the closed-loop process with the attack-sensitive parameters, $(K(t), L(t)) = (K_\Lambda, L_\Lambda)$.

The attack-sensitive parameters are chosen such that the attack-free closed-loop process operated with the attack-sensitive parameters is stable, and the process is destabilized by an attack. Particularly, the attack-sensitive parameters are chosen so that some eigenvalues of the augmented system matrix lie outside the unit circle (i.e., $\max_i |\lambda_i(A_\xi(\Lambda, K_\Lambda, L_\Lambda))| > 1$), and the matrix pair $(A_\xi(\Lambda, K_\Lambda, L_\Lambda), A_r(\Lambda))$ is observable. The attack-sensitive parameters are chosen to be sensitive to a range of attack magnitudes. Ideally, the attack-sensitive parameters may be chosen so that the range of attack magnitudes is as large as possible. The attack-sensitive parameters exploit the dependence of the terminal residual set on the control system parameters.

The switching instance $t_s$ and the period $T_c$ (called the cycle time) are the two design parameters for the proposed active detection methodology. The switching instance may be selected by a process operator based on operational considerations. For example, one way the switching instance may be selected is when the closed-loop performance degradation due to operation with attack-sensitive parameters is acceptable based on process economic considerations. The cycle time $T_c$ may be selected to balance a potential trade-off between attack detection and closed-loop performance and safety considerations. Given that $T_c$ is finite, the closed-loop augmented process state will remain bounded over the period when the attack-sensitive parameters are used in the control system ($t_s$ to $t_s + T_c$). This is true even if the closed-loop process is subjected to a destabilizing multiplicative attack during this period. Furthermore, it is expected that the likelihood of detecting potentially detectable and detectable attacks scales with $T_c$. Rigorous evaluation of this expectation is beyond the scope of the present section. However, the attack-sensitive parameters could result in closed-loop performance deterioration for the attack-free process compared to performance under the nominal control system parameters. Operating with the attack-sensitive parameters for long periods may not be desirable from a closed-loop performance perspective. Furthermore, for destabilizing attacks on the process operated under either control system (i.e., with nominal parameters or with the attack-sensitive parameters), the bound on the process state scales with $T_c$. If there is a state-space set whereby the process is operated safely, then $T_c$ should be selected to be small enough to ensure that the state is maintained within the safe set in the presence of a destabilizing attack. The closed-loop performance and safety considerations limit how long the cycle time should be.



Choose $K$, $L$ for screening against attack $\Lambda$
$\max|\lambda(A_{\xi_a})| < 1$
Yes
No
Attack is detectable
Compute:
1. $D_{\xi}^{est}$, $D_{\xi_a}^{est}$
2. $D_{r}^{est}$, $D_{r_a}^{est}$, $D_{r_e}^{est}$, $D_{r_{a,e}}^{est}$
3. $R^{est}$, $R_{a}^{est}$, $R_{e}^{est}$, $R_{a,e}^{est}$
$R_a^{est} \leq R_e^{est}$
Yes
Attack is undetectable
No
$R^{est} \leq R_{a,e}^{est}$
Yes
Attack is potentially detectable
No
The test is inconclusive

Fig. 3.2: Flowchart for the active attack detection methodology.

The main benefit of the proposed approach is to enhance the attack detection capabilities. An attacker may select a multiplicative attack that destabilizes the closed-loop process under the nominal parameters. Such an attack may be detected by the detection scheme in Eq. 3.10. If the attack is undetectable with the nominal control system parameters, it will not be detected. Even if the attack is potentially detectable, it may go undetected by the detection scheme. The active detection methodology enables the detection of attacks that are designed to be potentially detectable or undetectable with respect to the closed-loop process under nominal parameters.

Fig. 3.2 illustrates the flowchart for the active attack detection methodology. The proposed active detection methodology is summarized by the algorithm below. The algorithm is initialized with $t = 0$. The parameters of the methodology are the switching instance $t_s$ and the cycle time $T_c$.

1. If $t \in (t_s, t_s + T_c]$, set $(K(t), L(t)) = (K_\Lambda, L_\Lambda)$. Else, set $(K(t), L(t)) = (K^*, L^*)$.

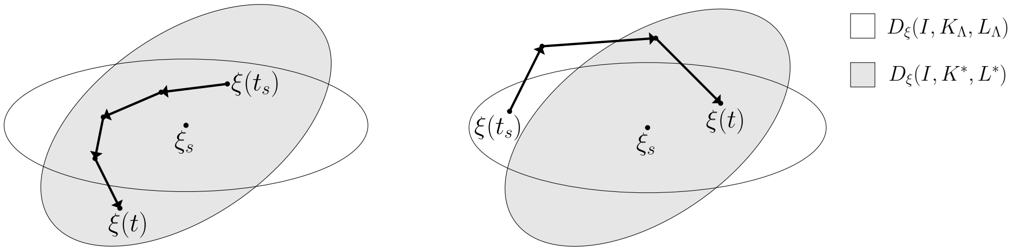
2. Compute the residual $r(t)$ and the output of the detection scheme in Eq. 3.10

3. If $z(t) = 1$, an attack is detected; implement attack identification and mitigation strategies. Else, an attack is not detected; go to Step 4.

4. Set $t \leftarrow t + 1$. Go to Step 1.

The methodology presented here illustrates a single switching cycle from the nominal to attack-sensitive parameters and back to the nominal parameters. To further enhance the detection capabilities, the methodology may be modified to include periodic switching from nominal to attack-sensitive parameters. Additionally, to detect a wider range of attack magnitudes, the methodology could be modified to include multiple control system switches from nominal parameters to other attack-sensitive parameters.

Under attack-free operation and when control system parameter switching takes place, the augmented state needs to be in the minimum invariant set of the attack-free process with the updated controller. In this way, the state moves from one minimum invariant set under one set of control system parameters to another. An example trajectory is shown in Fig. 3.3a. In this case, the control system parameters switch from nominal to attack-sensitive parameters when the augmented state is within the intersection of the minimum invariant sets with the nominal and attack-sensitive parameters, i.e., $\xi(t_s) \in D_\xi(I, K^*, L^*) \cap D_\xi(I, K_\Lambda, L_\Lambda)$. After the switch, the augmented state moves from the minimum invariant set with nominal parameters to the minimum invariant set with attack-sensitive parameters. However, the augmented state is not measured. When the control parameters switch, the augmented state may be outside the minimum invariant set of the process under the updated controller. If the process is attack-free, the augmented state will converge to the minimum invariant set, but the residual during the transient period may take values outside the terminal residual set, triggering a false alarm. An example of this is shown in Fig. 3.3b. For the attack-free process, the control system switches from nominal to attack-sensitive parameters at the time instance when the augmented state is outside the minimum invariant set associated with the attack-sensitive parameters, i.e., $\xi(t_s) \in D_\xi(I, K^*, L^*) \setminus D_\xi(I, K_\Lambda, L_\Lambda)$. As a result, the process dynamics is excited due to the switch, in the sense that the augmented state evolves briefly outside the minimum invariant set with attack-sensitive parameters. The residual during this transient period, may evolve outside the terminal residual set associated with the attack-sensitive parameters and trigger false alarms by the detection scheme in Eq. 3.11.



(a)

(b)

Fig. 3.3: (a) An example showing the evolution of the augmented state trajectory for the attack-free closed-loop process with a control parameter switch generating zero false alarms. (b) An example showing the evolution of the augmented state trajectory for the attack-free closed-loop process with a control system parameter switch that may generate false alarms.

Under an attack, a control system parameter switch may result in the augmented state exhibiting a transient behavior that mimics the transient behavior of the attack-free process. An example is illustrated in Fig. 3.4. After a control system parameter switch to attack-sensitive parameters occurs at time $t_s$, the augmented state of the attacked closed-loop process evolves outside its attack-free minimum invariant set until time $t_i$. However, after time $t_i$, the augmented state evolves within the attack-free minimum invariant set until the time instance $t_e$. This may result in the residuals of the process evolving briefly outside the attack-free terminal residual set, before converging to it. In this case, the alarms generated by the detection scheme monitoring the attacked process may be indistinguishable from the false alarm rate in the attack-free process.

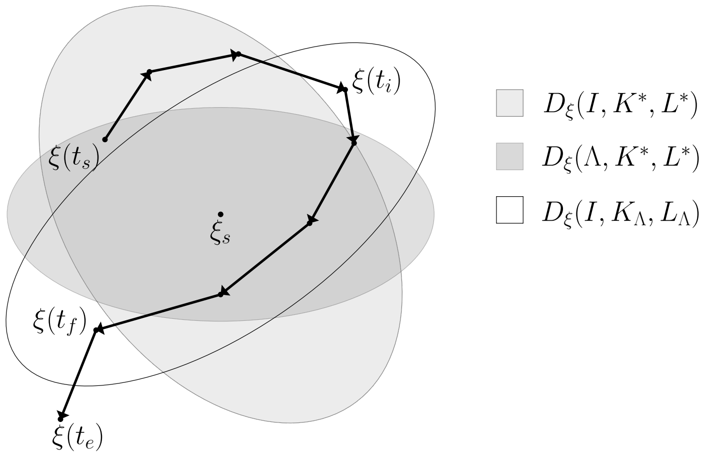


Fig. 3.4: An example showing the evolution of the augmented state trajectory for the attacked closed-loop process. After a control system parameter switch to attack-sensitive parameters, the augmented state mimics the transient behavior of an attack-free process briefly.

Owing to the complications described above, false alarms are not desirable. To minimize false alarms, a modification to the detection scheme in Eq. 3.11 may be considered. In particular, the detection scheme may be modified to generate an alarm only if the residual remains outside the terminal set for a specified period. The period may be chosen to span a few sample times to account for the potential transient behavior in the attack-free process. Using a timer threshold whereby the detection logic must deem abnormal operating behavior over a period before raising an alarm is a common approach for minimizing nuisance alarms [88].

**Remark 3.2.1.** *If an attack on the closed-loop process operated under either control mode, i.e., with nominal parameters or with the attack-sensitive parameters, is detected at any time $t_d \geq 0$, then attack identification and mitigation strategies could be employed to cope with the attack. These strategies are beyond the scope of this section and the subject of future work.*

**Remark 3.2.2.** *In addition to attack detection, the operating goals for a closed-loop process may be included as a constraint for selecting the attack-sensitive parameters. For* *example, it may be desired that the product concentration is within a certain range to ensure that the product is within specification. The attack-sensitive parameters may be selected to ensure that the potential values of the concentration in the corresponding minimum invariant set are within the acceptable range.*

**Remark 3.2.3.** *An attacker with prior knowledge of the active detection methodology may attempt to evade detection by using a destabilizing attack to target the process during the transient period after the control system switches from nominal parameters to attack-sensitive parameters. Randomly selecting the switching time ($t_s$) to minimize the possibility that the attacker knows when the controller switch occurs may be helpful in preventing the success of such attacks.*

**Remark 3.2.4.** *Detection of attacks that are potentially detectable under the nominal parameters is possible. In such cases, the attack identification and mitigation strategies could be activated following the detection of an attack while the closed-loop process is operated with the nominal parameters, and switching to the attack-sensitive parameters may not be needed.*

**Remark 3.2.5.** *Zero false alarms resulting from a parameter switch may be guaranteed under a special case when the minimum invariant set of the attack-free process under the updated parameters is a subset of the minimum invariant set of the process under the parameters used prior to the switch. For example, if the minimum invariant set for the attack-free process with nominal parameters is contained within the minimum invariant set for the process with attack-sensitive parameters (i.e., $D_\xi(I, K^*, L^*) \subset D_\xi(I, K_\Lambda, L_\Lambda)$), then, for the switch from nominal to attack-sensitive parameters, the augmented state of the attack-free process is contained within the minimum invariant set with attack-sensitive parameters, i.e., $\xi(t_s) \in D_\xi(I, K_\Lambda, L_\Lambda)$. As a result, there will be no transients, and zero false alarms can be guaranteed. However, zero false alarms cannot be guaranteed when a switch from the attack-sensitive parameters to the nominal parameters takes place because at the switching instance the augmented state may be outside the minimum invariant set associated with the nominal parameters, i.e., $\xi(t_s + T_c) \in D_\xi(I, K_\Lambda, L_\Lambda) \setminus D_\xi(I, K^*, L^*)$.*

*As a result, the second switch may generate false alarms by the detection scheme, and the detection scheme may need some modification to minimize false alarms.*

**Remark 3.2.6.** *The detectability of an attack is defined based on the ability of the residual-based detection scheme to detect the attack in finite time. It is a system property and is not influenced by the control parameter switching instance $t_s$ or the cycle time $T_c$. Under attack-free operation, both parameters ($t_s$ and $T_c$) do not influence closed-loop stability. The switching instance $t_s$ also does not influence closed-loop stability of the attacked process with either set of control system parameters, i.e., with nominal parameters or with attack-sensitive parameters.*

**Remark 3.2.7.** *Attacks on industrial control systems may take several forms. To characterize different types of attacks, the taxonomy of attacks on ICSs has been analyzed and presented in the literature [11, 12, 16–18]. In this section, the active detection methodology is designed to enhance the detection capabilities of a residual-based passive detection scheme, monitoring the process for multiplicative sensor-controller link cyberattacks. Multiplicative sensor-controller link attacks multiply the data communicated over the sensor-controller communication channels by a factor. Under a multiplicative attack, the real-time process operational data communicated over the communication channels is masked by the attack. Replay attacks communicate historic attack-free process operational data over the compromised controller communication channels and are fundamentally different from multiplicative attacks. Under a replay attack, the data communicated over the compromised controller communication channels has no correlation to the real-time process operational data. Characterization of the detection capability of the proposed active detection methodology to detect other types of cyberattacks is beyond the scope of the work presented here.*

**Remark 3.2.8.** *The proposed active detection methodology considers a single switch from nominal mode (during which the process is operated with nominal parameters) to the attack-sensitive mode (during which the process is operated with attack-sensitive parameters) and back to operating in the nominal mode thereafter. To further enhance the detection capabilities of the passive residual-based detection scheme with respect to a wider range* *of attack magnitudes, the proposed methodology may be modified to include control system switches between the nominal mode and multiple attack-sensitive modes. The closed-loop process with the active detection methodology using multiple control system switches may be considered a switched system. Successive switching between different modes may compromise the closed-loop stability of the process. Here, closed-loop stability for the switched system under bounded process disturbances and measurement noise means ultimate boundedness of the process state (and estimation error) in a small neighborhood of the origin. No guarantees can be made on the closed-loop stability of the attacked process without placing limitations on the attack magnitude. To guarantee the closed-loop stability of the attack-free process, two classical approaches may be employed (e.g., [73]). In the first approach, a common Lyapunov function may be used to find the control parameters for the nominal mode and attack-sensitive modes. The advantage of this approach is that the Lyapunov function value will decrease over time under any mode if the state is sufficiently far from the origin. However, this approach restricts the choice of control parameters. As an alternative approach, a Lyapunov function may be derived for each mode, i.e., the multiple Lyapunov function approach. In this case, the switching times must be carefully selected because the Lyapunov function value of the inactive modes may increase over time when another mode is active. Nonetheless, existing methods for determining the switching times could be employed (see, for example, [73]).*

## 3.3 Application to a Chemical Process

A chemical process consisting of a CSTR is considered where a second-order, exothermic reaction of the form $A \to B$ occurs. The CSTR contents are assumed to be well-mixed, and the contents may be heated or cooled using, for example, a cooling jacket or submerged heat exchanger coil. A dynamic process model is obtained from mass and energy balances under standard modeling assumptions, and is given by the following system of ordinary differential equations:

$$\begin{aligned} \frac{dC_A}{dt} &= \frac{F}{V}(C_{A0} + \Delta C_{A_0} - C_A) - k_0 e^{\frac{-E}{RT}} C_A^2 \\ \frac{dT}{dt} &= \frac{F}{V}(T_0 + \Delta T_0 - T) - \frac{\Delta H k_0}{\rho C_p} e^{\frac{-E}{RT}} C_A^2 + \frac{Q}{\rho C_p V} \end{aligned} \tag{3.12}$$

Parameters for the CSTR

| | |
|---|---|
| Volumetric flow rate ($F$) | $5.0\,\mathrm{m^3\,h^{-1}}$ |
| Reactor volume ($V$) | $1.0\,\mathrm{m^3}$ |
| Feed concentration of $A$ ($C_{A0}$) | $4.0\,\mathrm{kmol\,m^{-3}}$ |
| Activation energy ($E$) | $5.0 \times 10^4\,\mathrm{kJ\,kmol^{-1}}$ |
| Pre-exponential factor ($k_0$) | $8.46 \times 10^6\,\mathrm{m^3\,h^{-1}\,kmol^{-1}}$ |
| Gas constant ($R$) | $8.314\,\mathrm{kJ\,kmol^{-1}\,K}$ |
| Feed temperature ($T_0$) | $300\,\mathrm{K}$ |
| Density of reactor liquid hold-up ($\rho$) | $1000\,\mathrm{kg\,m^{-3}}$ |
| Heat of reaction ($\Delta H$) | $-1.15 \times 10^4\,\mathrm{kJ\,kmol^{-1}}$ |
| Heat capacity ($C_p$) | $0.231\,\mathrm{kJ\,kg\,K^{-1}}$ |
| Steady-state heat rate added/removed from the reactor ($Q_s$) | $0\,\mathrm{kJ\,h^{-1}}$ |
| Steady-state reactant concentration ($C_{As}$) | $1.22\,\mathrm{kmol\,m^3}$ |
| Steady-state temperature ($T_s$) | $438.2\,\mathrm{K}$ |

where $C_{A0}$ is the feedstock reactant concentration, $T_0$ is the feedstock temperature, $C_A$ is the reactor reactant concentration, $T$ is the reactor temperature, and $Q$ is the heat added to or removed from the tank contents. The variables $\Delta C_{A_0}$ and $\Delta T_0$ represent two bounded process disturbances, modeled as a deviation from the nominal feedstock reactant concentration and temperature. The process parameter definitions and their values are given in Table 2.1, and reproduced in this chapter to make it self-contained. The control objective is to operate the process around its open-loop stable steady-state with $C_A = C_{As}$ and $T = T_s$ where the values are given in Table 2.1. The measured outputs are $C_A$ and $T$, and the manipulated input is $Q$. Defining deviation variables, the state, input, process disturbance, and output are given by: $x = [x_1\ x_2]^T = [C_A - C_{As}\ T - T_s]^T$, $u = Q - Q_s$, $w = [\Delta C_{A0}\ \Delta T_0]^T$, and $y = [x_1 - x_{1s}\ x_2 - x_{2s}]^T$. The sensors measuring $C_A$ and $T$ are corrupted by bounded measurement noise.

To design a linear feedback control law for the CSTR, the nonlinear process model in Eq. 3.12 is linearized about its steady-state, yielding a continuous-time linear process model. The continuous-time linear model is discretized in time with a sampling interval of $10^{-2}$ h by assuming a zeroth-order hold of the inputs to obtain a discrete-time linear process model of the form in Eq. 3.1. The system matrices are given by:

$$A = \begin{bmatrix} 0.7364 & -0.0041 \\ 10.6953 & 1.1560 \end{bmatrix},\ B = \begin{bmatrix} -9.0708 \times 10^{-8} \\ 4.6741 \times 10^{-5} \end{bmatrix},\ G = \begin{bmatrix} 0.0433 & -0.0001 \\ 0.2724 & 0.0540 \end{bmatrix} \tag{3.13}$$

The discrete-time linear model for the CSTR of the form in Eq. 3.1 with matrices in Eq. 3.13 is referred to the linearized CSTR model.

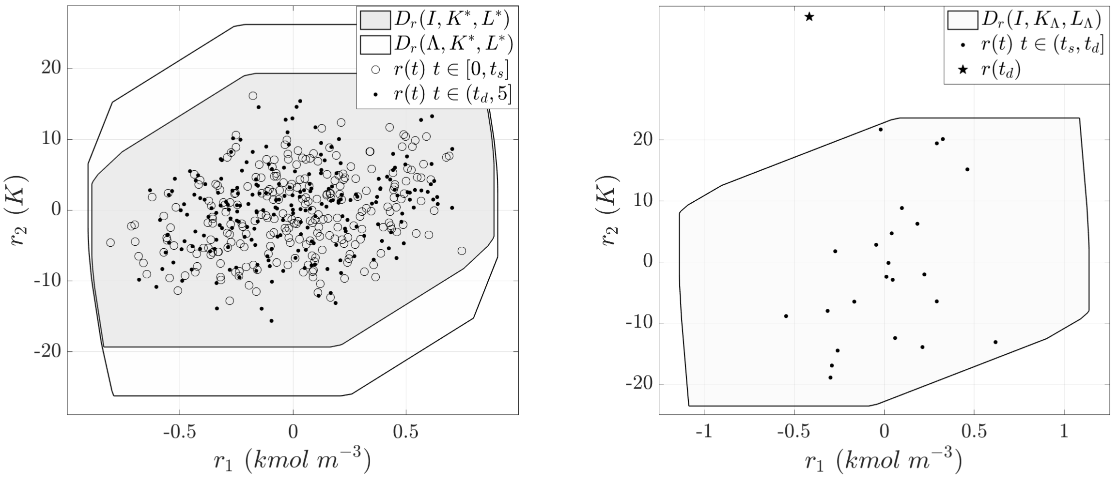
In the simulations presented in subsequent sections, the Multi-Parametric Toolbox (MPT) 3.0 [82] is used for the calculation of the minimum invariant and residual terminal sets. Numerical approximations of the minimum invariant sets are computed based on the algorithm in [79] with an error bound of $5 \times 10^{-5}$. In comparing the numerical estimates of the terminal residual sets, the technique presented in [22] is used. In the remainder, time is represented in continuous time with a slight abuse of notation. In Section 3.3.1, the application of the active detection methodology for enhancing the detection capabilities of the residual-based detection scheme is demonstrated using the linearized CSTR model. In Section 3.3.2, the active detection methodology is applied to the nonlinear CSTR to evaluate the efficacy of the proposed approach when dealing with more complex process dynamics.

### 3.3.1 Application of the Active Detection Methodology to the Linearized CSTR

In this section, the CSTR is modeled using the linearized process model. The disturbance set ($F$) is described by an admissible process disturbance set ($W$) given by $\Delta C_{A0} \in [-0.5, 0.5]$ kmol m$^{-3}$ and $\Delta T_0 \in [-5, 5]$ K, and an admissible measurement noise set ($V$) described by $[-0.5, 0.5]$ kmol m$^{-3}$ and $[-5, 5]$ K for the concentration and temperature sensors, respectively. The control actions are computed using a linear control law of the form in Eq. 3.4 using estimates generated by a Luenberger observer of the form in Eq. 3.3. Pole placement is used to determine the controller and observer gains. The nominal parameters $(K^*, L^*)$ are chosen to stabilize the attack-free process with the controller gain computed with poles placed at $[0.2\ -0.1]$, and the observer gain with the poles placed at [0.2 0.3]. For the attack-sensitive parameters $(K_\Lambda, L_\Lambda)$, the controller gain is computed with poles placed at $[-0.33 \ \ -0.3]$ and the observer gain is computed with poles placed at $[-0.2 \ \ -0.3]$. In the absence of an attack, the closed-loop process with attack-sensitive parameters is stable in the sense that $\max_i |\lambda_i(A_\xi(I, K_\Lambda, L_\Lambda))| = 0.33 < 1$. Furthermore, the attack-sensitive parameters are found to be "sensitive" to multiplicative sensor-controller link attacks with magnitudes in the set $\{\Lambda \mid \text{diag}(1, \alpha) \mid \alpha \in [0.6, 0.95]\}$. This range is numerically verified by parameterizing the attack magnitude with a parameter $\alpha$ where $\Lambda = \text{diag}(1, \alpha)$. The value of $\alpha$ is varied beginning at 0.6 and incremented by 0.01 until a value of $\alpha = 0.95$ is reached. For each $\Lambda$, the control system parameters are sensitive to the attack if any of the eigenvalues of $A_\xi(\Lambda, K_\Lambda, L_\Lambda)$ are outside the unit circle and the observability matrix for the pair $(A_\xi(\Lambda, K_\Lambda, L_\Lambda), A_r(\Lambda))$ is full rank. A similar analysis is performed using the nominal parameters. The nominal parameters are not sensitive to any attack in the set $\{\Lambda \mid \text{diag}(1, \alpha) \mid \alpha \in [0.6, 0.95]\}$.

Three sets of simulations are performed, and the results are compared. First, the closed-loop process with nominal parameters and without the active detection methodology (without switching) is considered. Second, the active detection methodology is applied to the attacked closed-loop process. The first and second simulation sets are used to evaluate the enhanced detection capabilities of the proposed active detection methodology. Third, the active detection methodology is applied to the attack-free process to identify if false alarms are raised resulting from control system parameter switching.

Each simulation set consists of 1000 simulations of the closed-loop process. The process disturbances and measurement noise are modeled as random variables drawn from a uniform distribution on the interval defined by the bounds of the appropriate admissible set. The value of the random variables modeling the process disturbances and measurement noise are varied every sample time, and different realizations of the random variables are used in each simulation. The same realizations of random variables are used across simulation sets to compare the results across simulation sets. For each simulation, the process states are initialized at 0, and a period of 5 h is simulated.



(a)

(b)

Fig. 3.5: The residual values of the attacked linearized CSTR process with (a) with nominal parameters before and after a switch to the attack-sensitive parameters and (b) with attack-sensitive parameters where the attack is detected at time $t_d = 2.74\,\mathrm{h}$.

For simulating the attacked process, an attack magnitude of $\Lambda = \mathrm{diag}(1, 0.9)$ is considered. Under the nominal parameters, the attack is potentially detectable, which can be observed from Fig. 3.5a since $D_r(\Lambda, K^*, L^*) \nsubseteq D_r(I, K^*, L^*)$. Under the attack-sensitive parameters, the attack is detectable, and the terminal residual set for the attack-free process under the attack-sensitive parameters is shown in Fig. 3.5b. To demonstrate the enhancement of attack detection capabilities of the residual-based detection scheme, the proposed active detection methodology is applied and the control system switches to attack-sensitive parameters at time $t_s = 2.5\,\mathrm{h}$. A cycle time of $T_c = 1\,\mathrm{h}$ is used, i.e., in the absence of attack detection, a second switch from attack-sensitive to nominal parameters is implemented at time $t_s + T_c = 3.5\,\mathrm{h}$.

The residual values for one of the simulations from the first set (with the active detection methodology) are depicted in Fig. 3.5. From Fig. 3.5a, the residual values of the closed-loop process with nominal parameters are in the attack-free terminal residual set before the switch occurs, i.e., $r(t) \in D_r(I, K^*, L^*)$, $t \in [0, 2.5]$ h. As a result, no alarms are raised by the detection scheme, and the attack is not detected during this period. After the switch to attack-sensitive parameters, the attack is detected at time $t_d = 2.74\,\mathrm{h}$

because the residual value is outside the terminal residual set, i.e., $r(t_d) \notin D_r(I, K_\Lambda, L_\Lambda)$ (Fig. 3.5b). Following the detection of the attack, the control system switches back to nominal parameters to stabilize the closed-loop process. In practice, attack identification and mitigation measures would be activated after detection. After switching back to the nominal parameters, no alarms are raised (Fig. 3.5a).

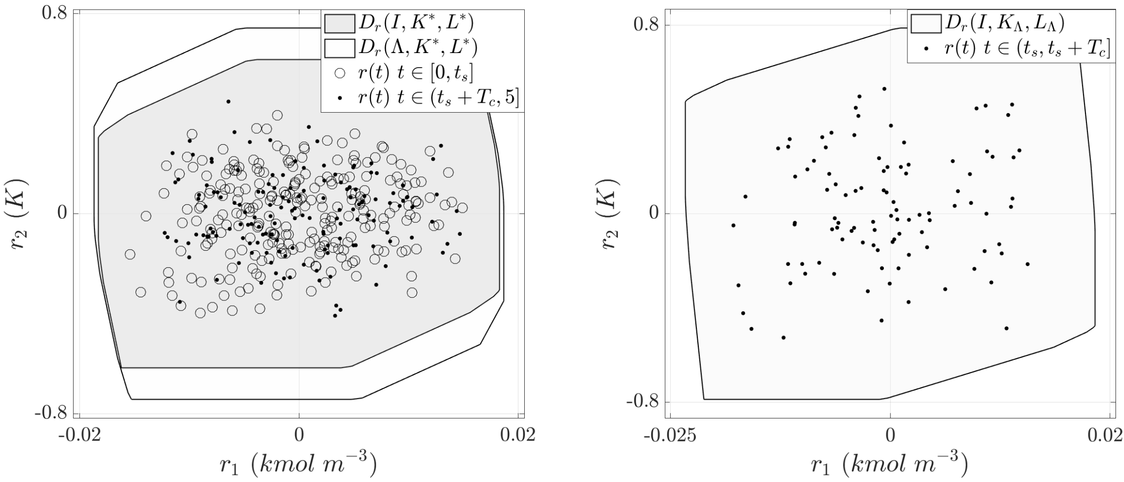
The results obtained from the first and second simulation sets are compared. For the second simulation set (passive detection), the attack is detected in 43 out of 1000 simulations. In 19 of these 43 simulations, the attack is detected before the switching time ($t_s = 2.5\,\mathrm{h}$). For the first and second simulation sets, the process evolves the same during the period $t = 0\,\mathrm{h}$ to $t = 2.5\,\mathrm{h}$ because the same control system, attack, disturbance, measurement noise, and detection scheme are applied to the process during this period. For the 19 simulations, switching to attack-sensitive parameters is not needed as the attack is detected before the switch. For the remaining 981 simulations, the attack is detected after switching to the attack-sensitive parameters within a maximum of 24 sample times in all simulations. Considering the 981 remaining simulations with passive detection (second simulation set), the attack is detected in 33 sample times after $t_s = 2.5\,\mathrm{h}$ in the best case and never detected over the simulated $5\,\mathrm{h}$ operating period in the worst case. The results demonstrate an enhancement of detection capabilities of the residual-based detection scheme by applying the active detection methodology.

In the third simulation set, the false alarm rate under the proposed active detection methodology is evaluated. The attack-free process is considered for the analysis. False alarms are not raised after switching into and out of the attack-sensitive parameters in any simulation. Further analysis is performed to address the possibility of false alarms. The containment of the augmented state at the switching instances in the minimum invariant set is verified. When switching to the attack-sensitive parameters, the state is verified to be in the minimum invariant set associated with the attack-sensitive parameters, i.e., $\xi(t_s) \in D_\xi(I, K_\Lambda, L_\Lambda)$. When switching back to the nominal parameters, the state is verified to be in the minimum invariant set associated with the nominal parameters, i.e., $\xi(t_s + T_c) \in D_\xi(I, K^*, L^*)$. When the control system switches from nominal parameters to attack-sensitive parameters, the augmented state is in the minimum invariant set of the attack-free process with attack-sensitive parameters in all simulations. When the control system switches from attack-sensitive parameters to nominal parameters, the augmented state is not contained within the minimum invariant set of the attack-free process with nominal parameters in 958 out of the 1000 simulations. In the 958 simulations, the augmented state evolves briefly outside the minimum invariant set of the attack-free closed-loop process with nominal parameters. The augmented state converges to the minimum invariant set within two sample times. No false alarms are observed in any of these cases. Although this analysis confirms the possibility of a false alarm, false alarms are not raised in these cases.

### 3.3.2 Application of the Active Detection Methodology to the Nonlinear CSTR

In this section, the active detection methodology is applied to the nonlinear CSTR process model in Eq. 3.12. To this end, the state is maintained within a region around the origin when the process disturbances and measurement noise are small. In this region, the nonlinear process may be approximated by its linearized model. As the magnitude of the disturbances and measurement noise increases, the impact of the nonlinearities increases. While the theoretical results in this section are developed strictly for linear systems, the objective of this study is to assess the method's applicability to the nonlinear case. The proposed active detection method is therefore applied to the nonlinear process, considering small disturbances. The disturbance set $F$ is described by an admissible process disturbance set ($W$) given by $\Delta C_{A0} \in [-0.01, 0.01]$ $\mathrm{kmol\,m^{-3}}$ and $\Delta T_0 \in [-0.2, 0.2]$ K, and an admissible measurement noise set ($V$) described by $[-0.01, 0.01]$ $\mathrm{kmol\,m^{-3}}$ and $[-0.2, 0.2]$ K for the concentration and temperature sensors, respectively. The process disturbances and measurement noise are modeled as random variables drawn from a uniform distribution in the interval specified by the bounds of the admissible set. The same realizations of random variables are used across simulation sets.

The closed-loop simulations of the continuous-time CSTR process use the explicit Euler's method with a step size of $1 \times 10^{-4}\,\mathrm{h}$ to integrate the ordinary differential equations in Eq. 3.12. Extensive simulations are employed to verify that further reduction in the integration time step did not lead to substantial changes in the computed solution of the nonlinear ordinary differential equations. To steer the process states to the origin, a linear control law (Eq. 3.4) is used with state estimates generated by a Luenberger observer (Eq. 3.3) based on the linearized process model using the matrices $A$, $B$, and $G$ (Eq. 3.13). The sampling period of the control system is $10^{-2}$ h, which is the same as that used in Section 3.3.1. In general, the sampling period should be sufficiently small so that the continuous-time process in Eq. 3.12 may be stabilized with the discrete-time controller in Eq. 3.4. Two sets of simulations, each consisting of 1000 simulations of the attack-free closed-loop process with nominal parameters and the attack-free closed-loop process with attack-sensitive parameters are performed. In all simulations, the attack-free closed-loop process is found to be stable, verifying that the sampling period is appropriately chosen. Numerical approximations of the attack-free terminal residual sets for the closed-loop process with nominal parameters and with the attack-sensitive parameters are computed from the linearized CSTR model. To verify that the terminal residual set approximated from the linear model is a suitable approximation for the nonlinear process, the evolution of the residuals are considered under the nominal and attack-sensitive parameters. Considering the same two sets of simulations used for verifying closed-loop stability, the residuals of the attack-free process are bounded within the appropriate terminal residual set in all simulations. Based on this, the computed terminal residual sets are suitable approximations for the nonlinear process.

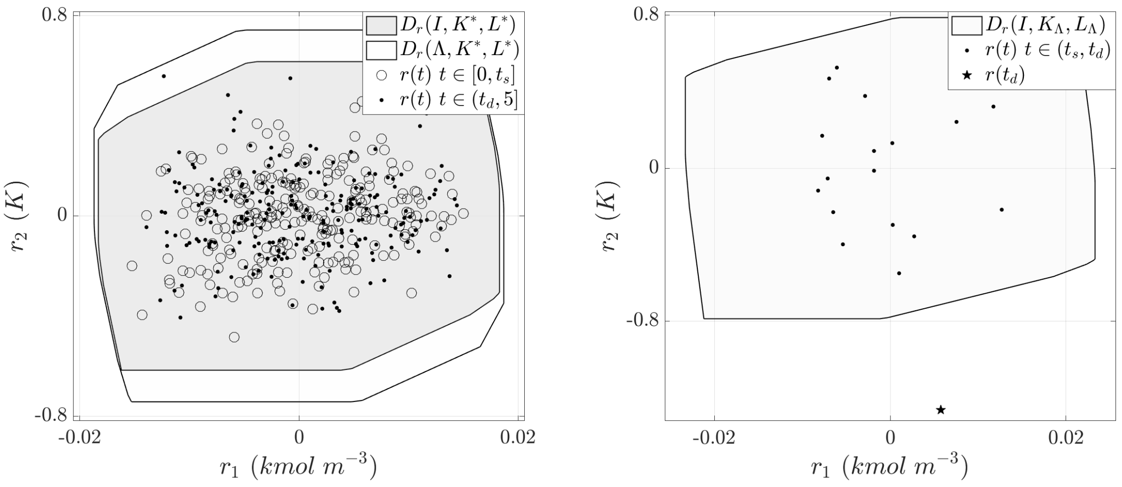


(a)

(b)

Fig. 3.6: (a) The residual values of the attack-free nonlinear CSTR process with nominal parameters before and after a switch to the attack-sensitive parameters. (b) The residual values of the attack-free nonlinear CSTR process with attack-sensitive parameters.

A similar study as that performed in Section 3.3.1 consisting of three simulations sets is carried out for the nonlinear process. In each simulation set, 1000 simulations are conducted. These simulations enable the evaluation of the detection capabilities and potential of false alarms under the proposed active detection methodology for the nonlinear process. For the first two simulation sets, an attack of magnitude $\Lambda = \text{diag}(1, 0.9)$ is considered. Based on the linearized model, the attack-sensitive parameters are "sensitive" to this attack, while the nominal parameters are not. For the third simulation set, attack-free operation is considered. In the first simulation set (with active detection methodology), the control system switches from nominal parameters to attack-sensitive parameters at time instance $t_s = 2.5\,\text{h}$. The cycle time under the attack-sensitive parameters is $T_c = 1\,\text{h}$. The residual values from one simulation in the first simulation set (with active detection methodology) are depicted in Fig. 3.7. From Fig. 3.7b, the attack is detected at time $t_d = 2.69\,\text{h}$ after the control system switches from nominal parameters to attack-sensitive. Following the detection of the attack, the control system switches back to nominal parameters to stabilize the closed-loop process. From Fig. 3.7a, the residual is outside the attack-free terminal residual set for one sample time after switching back to nominal parameters. As a result, another alarm is raised. Thereafter, no alarms are raised because the residual converges to the terminal residual set.



(a) (b)

Fig. 3.7: (a) The residual values of the attacked nonlinear CSTR process with nominal parameters before and after a switch to the attack-sensitive parameters. (b) The detection of the attack at the time $t_d = 2.69\,\text{h}$ after a switch from nominal parameters to attack-sensitive parameters.

For the 1000 simulations of the attacked process under nominal parameters monitored by the residual-based detection scheme in Eq. 3.10 (passive only detection), the attack is detected in only one simulation. In the active detection simulation set, the attack is detected after the control system switches from nominal parameters to attack-sensitive parameters in all cases. After switching to attack-sensitive parameters, the attack is detected within a minimum of 2 sample times and a maximum of 19 sample times after the switching instance. Thus, the active detection methodology also enhances the detection capabilities of the residual-based detection scheme for the nonlinear process.

In the third simulation set, 1000 simulations of the attack-free process with the active detection methodology are performed to analyze the false alarm rate. No false alarms are raised in any of the simulations. Similar to the analysis for the third simulation set in Section 3.3.1, the containment of the augmented state at the switching instance within the minimum invariant set with the updated parameters is verified. At the switching instance $t_s$, the state is always contained in the minimum invariant set $D_\xi(\Lambda, K^*, L^*)$

in all cases. At the switching instance $t_s + T_c$, the augmented state is not within the minimum invariant set under nominal parameters in 921 out of 1000 simulations. Over these 921 simulations, the augmented state evolves outside the minimum invariant set, but converges to it in 2 sample times. However, no false alarms are observed in any of the 921 simulations.

#### 3.3.2.1 Comparison between Active and Passive Detection

The application of active detection methodology to enhance the detection capabilities of the CUSUM detection scheme, another residual-based detection scheme, is demonstrated. The CUSUM detection scheme is a statistical change detection scheme that monitors a process based on the deviation of the detection metric from a predefined baseline value. Application of the CUSUM detection scheme using the residual vector as the detection scheme has been considered as a passive attack detection scheme previously in the literature [21, 22, 38]. The CUSUM detection scheme monitoring a process based on the 2-norm of the residual may be represented by:

$$S(t) = S(t-1) + \|r(t)\| - b;\ S(-1) = 0; \tag{3.14}$$

where $S(t)$ is the CUSUM statistic, which is the detection scheme output, $r(t)$ is the residual of the process at the time $t \geq 0$, and $b$ is the baseline parameter. An attack on the process is detected by the scheme if the CUSUM statistic exceeds the tolerance value, which is the alarm threshold $\tau$, i.e.,

$$\begin{gathered} S(t) \leq \tau;\ \text{No Attack} \\ S(t) > \tau;\ \text{Attack} \end{gathered}$$

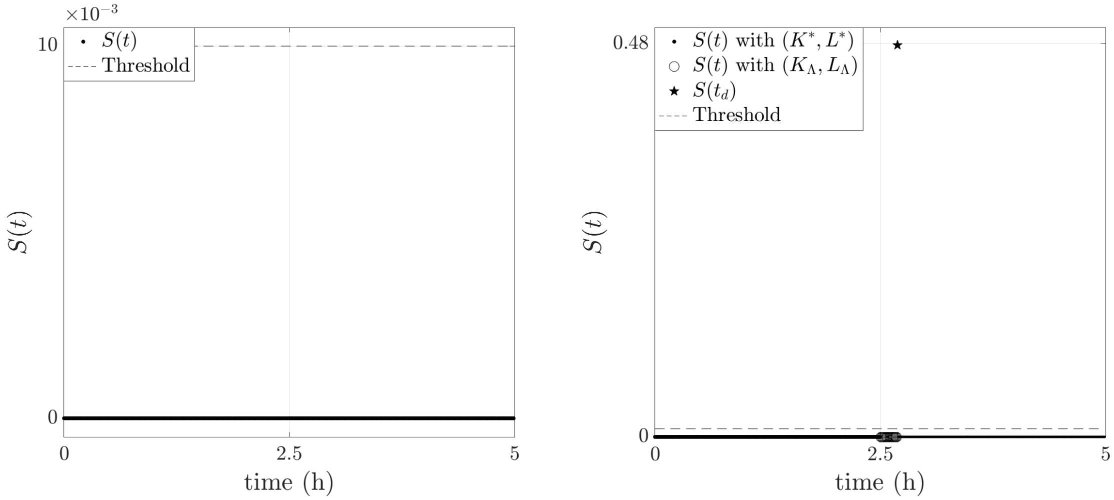
The CUSUM detection scheme is chosen as the detection scheme instead of a set membership-based detection scheme which was considered earlier to monitor the CSTR. An attack is considered in the temperature sensor-controller link and has the same magnitude as that previously considered, i.e., $\Lambda = \text{diag}(1, 0.9)$. To tune the CUSUM detection scheme for a zero false alarm rate in the absence of an attack (and without switching) when monitoring the closed-loop process, the approach presented in [22] is adopted. Because the residuals of the attack-free closed-loop process are always contained within the terminal residual set, they are also contained within the 2-norm ball enclosing the terminal residual set. Consequently, the norm of the residual vector of the attack-free process is always less than the radius of the ball. The baseline parameter is selected as the radius of the 2-norm ball enclosing the terminal residual set of the attack-free process. With the nominal parameters, the radius is $R_{D_r(I,K^*,L^*)} = 0.6169$, and with the attack-sensitive parameters, the radius is $R_{D_r(I,K_\Lambda,L_\Lambda)} = 0.7884$. Based on Eq. 3.14, the CUSUM statistic for the attack-free process always remains at zero with this choice of the baseline parameter, and any non-zero CUSUM statistic value may be considered indicative of an attack. In this case, the CUSUM detection may be tuned with an alarm threshold choice of $\tau = 0$. To maintain a zero false alarm rate when there are small variations in the process that are not necessarily due to an attack, the alarm threshold for the detection scheme is set at $\tau = 0.01$. Furthermore, the CUSUM detection scheme is implemented so that upon detection of an attack, the CUSUM statistic is reset to 0 at the next time step, i.e., $S(t) > \tau$ implies $S(t+1) = 0$.

To enhance the attack detection capability of the CUSUM detection scheme, the active detection methodology is implemented. Because the parameter $b$ is dependent on the terminal residual set, which depends on the control system parameters, the baseline parameter switches from the value with the nominal parameters to the value corresponding to the attack-sensitive parameters when operating with the attack-sensitive parameters:

$$b(t) = \begin{cases} R_{D_r(I,K_\Lambda,L_\Lambda)} = 0.7884; \; t \in (t_s, t_s + T_c] \\ R_{D_r(I,K^*,L^*)} = 0.6169; \; \text{Otherwise} \end{cases}$$

In this case, the control system switches back to using the nominal parameters if an attack is detected during operation with the attack-sensitive parameters.

Two sets of simulations are performed for the process monitored by the CUSUM detection scheme in Eq. 3.14. First, the attacked closed-loop process with nominal parameters (no parameter switching) and monitored by the CUSUM detection scheme is considered. Next, the attacked closed-loop process with the active detection methodology and monitored by the CUSUM detection scheme is considered. The attack is not detected in any of the simulations using passive detection (with no switching). Fig. 3.8a illustrates the output of the CUSUM scheme for one simulation with passive detection. With the active detection methodology, however, the attack is detected in all cases after the control system switches from nominal parameters to attack-sensitive parameters. The attack is detected in a minimum of 2 sample times and a maximum of 19 sample times after switching.

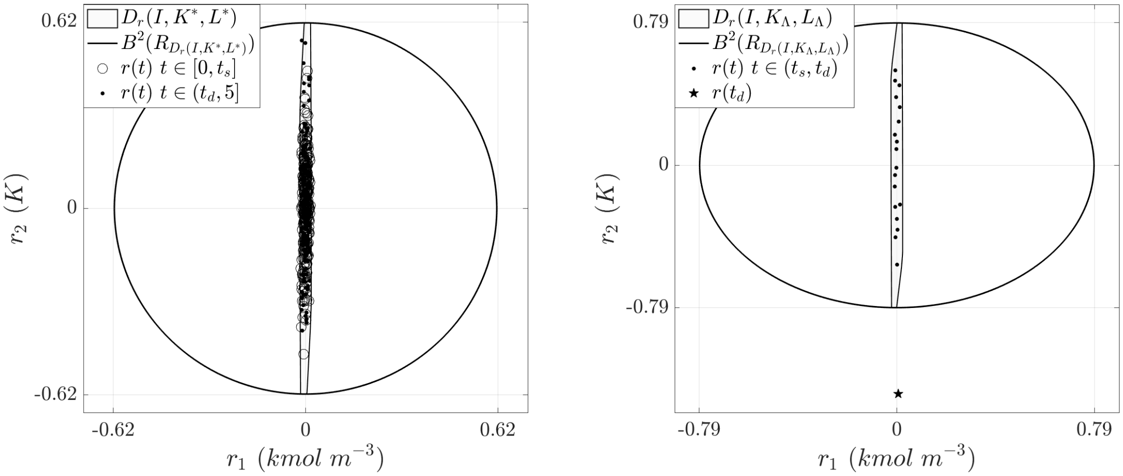


(a)

(b)

Fig. 3.8: (a) The CUSUM statistic for the attacked CSTR without the active detection methodology implemented. (b) The CUSUM statistic for the attacked CSTR process with the active detection methodology implemented, showing that the attack is detected at time $t = 2.69$ h.

The residual values from one simulation with the active detection methodology are shown in Fig. 3.9. Before the switch to attack-sensitive parameters occurs, the residual values for the attacked closed-loop process with nominal parameters are in the 2-norm ball enclosing the attack-free terminal residual set, i.e., $r(t) \in B^2(R_{D_r(I,K^*,L^*)})$ for all $t \in [0, 2.5]$ h (Fig. 3.9a). As a result, no alarms are raised by the detection scheme. After the switch to attack-sensitive parameters, the attack is detected at time $t_d = 2.69\,\text{h}$ because the residual value is outside the 2-norm ball enclosing the terminal residual set (Fig. 3.9b). From Fig. 3.8b, an alarm is raised by the detection scheme at the detection time. Following the detection of the attack, the control system switches back to nominal parameters to stabilize the closed-loop process. From Fig. 3.9a, no alarms are raised by the detection scheme thereafter.



(a) (b)

Fig. 3.9: (a) The residual values of the attacked nonlinear CSTR with nominal parameters before and after a switch to attack-sensitive parameters. (b) The residual values of the attacked nonlinear CSTR with attack-sensitive parameters showing attack detection at time $t_d = 2.69$ h.

**Remark 3.3.1.** *The CUSUM detection scheme is a dynamic detection scheme measuring the cumulative deviation of the 2-norm of the residual from the baseline parameter over time. While not observed in the simulations presented in this section, in some cases, the CUSUM detection scheme may not detect an attack immediately after the residual of the closed-loop process leaves the 2-norm ball enclosing its attack-free terminal residual set. However, the set membership-based detection scheme in Eq. 3.10 detects an attack as soon as the residual leaves the terminal residual set of the attack-free process. Based on this, it may appear that the CUSUM detection scheme is not as sensitive to the drifts in the detection parameter, as the set membership-based detection scheme. However, the sensitivity of the CUSUM detection scheme to the drifts in the detection metric is dependent on its tuning parameters (i.e., on the threshold $\tau$ and the parameter $b$). The tuning approach is fundamentally different from the set membership-based detection scheme. Consequently, the detection performance of the CUSUM detection scheme with a given choice of $\tau$ and $b$ may not be directly compared with that of the set membership-based detection scheme.*

## 3.4 Conclusions

In this chapter, an active attack detection methodology that enhances the attack detection capabilities of residual-based detection schemes was developed. The methodology utilizes control system parameter switching to probe for, and elicit detection of, multiplicative sensor-controller attacks. In this approach, the control system switches occasionally between nominal control system parameters, selected on the basis of standard control design criteria, and attack-sensitive control system parameters to manage the potential trade-off between closed-loop performance and attack detectability. The relationship between attack detectability with respect to a residual-based detection scheme, the control system parameters, and closed-loop stability was rigorously analyzed and the selection of the attack-sensitive parameters exploited this relationship. The enhancement of attack detection capabilities of two residual-based detection schemes upon application of the active attack detection methodology was demonstrated using a chemical process example.

# Chapter 4

# A Control-Switching Approach for Cyberattack Detection in Process Systems with Minimal False Alarms

In this chapter, an active detection method utilizing controller-observer gain switching with minimal false alarms is developed. As part of the proposed active detection method, the control system operates under two modes. Under the first mode, called the nominal mode, the control system operates with controller-observer parameters selected using traditional control design criteria. Under the second mode (the "attack-sensitive" mode), the control system operates with controller-observer parameters selected to enhance the detection capability of an output and residual-based detection scheme. The active detection method manages the trade-off between closed-loop performance and attack detectability. Since switching may excite the process dynamics, generating false alarms, a state-dependent switching condition that guarantees zero false alarms is developed using a region containing the attack-free process states, called the confidence region. Practical implementation issues related to the active detection method are discussed, including the inability to ensure that the switching condition will be satisfied over the time interval it is desired to switch the control system. Switching between the nominal and attack-sensitive modes may be desirable even if the switching condition is not satisfied. A modified active detection method for minimizing false alarms that incorporates the switching condition while balancing the practical requirement to switch between modes is proposed. The application of the proposed active detection method in attack detection and minimizing false alarms is demonstrated using two illustrative process examples.

## 4.1 Preliminaries

### 4.1.1 Notation and Definitions

For an $n$-dimensional vector $x \in \mathbb{R}^n$, $\|x\| := (\sum_{i=1}^{n} x_i^2)^{\frac{1}{2}}$ represents the Euclidean norm. For the compact set $X \subset \mathbb{R}^n$, $AX := \{Ax \mid x \in X\}$, where $A$ is a matrix. The Minkowski sum of two sets, $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^n$ is represented by $X \oplus Y = \{x + y \mid x \in X, y \in Y\}$. The Minkowski difference of two compact and convex sets, $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^n$ is represented by $X \ominus Y = \{x - y \mid x \in X, y \in Y\}$. For a square matrix $A$, $\lambda_i(A)$ represents the $i^{th}$ eigenvalue of $A$. $\text{diag}(\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_n)$ represents an $n \times n$ diagonal matrix with diagonal elements $\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_n$.

### 4.1.2 Class of Processes

In this work, discrete-time linear time-invariant processes are considered:

$$x(t+1) = Ax(t) + Bu(t) + Gw(t) \tag{4.1}$$

where $x(t) \in \mathbb{R}^{n_x}$ is the process state vector, $u(t) \in \mathbb{R}^{n_u}$ is the manipulated input vector, and $w(t) \in W \subset \mathbb{R}^{n_w}$ is the bounded process disturbance vector. The set $W$ is assumed to be known and described by a convex polytope containing the origin. The measured output $(y(t))$ is subject to measurement noise and may be altered by a multiplicative sensor-controller link attack:

$$y(t) = \Lambda(Cx(t) + v(t)) \tag{4.2}$$

where $y(t) \in \mathbb{R}^{n_x}$ is the measured output vector and $v(t) \in V \in \mathbb{R}^{n_x}$ is the bounded measurement noise vector. The set $V$ is assumed to be known and described by a convex polytope containing the origin. The matrix $C$ is assumed to be invertible. The matrix $\Lambda \in \mathbb{R}^{n_x \times n_x}$ is used to model the multiplicative sensor-controller link attack on the process and is called the attack magnitude. When $\Lambda = I$, the process is attack-free. Without loss of generality, the origin of the unforced process (Eq. 4.1 with $u \equiv 0$ and $d \equiv 0$) is assumed to the desired operating steady-state.

A Luenberger observer is synthesized to estimate the process states in Eqs. 4.1-4.2:

$$\hat{x}(t+1) = A\hat{x}(t) + Bu(t) + L(y(t) - \hat{y}(t)) \tag{4.3a}$$

$$\hat{y}(t) = C\hat{x}(t) \tag{4.3b}$$

where $\hat{x}(t) \in \mathbb{R}^{n_x}$ is the state estimate generated by the observer, $\hat{y}(t) \in \mathbb{R}^{n_x}$ is the estimated output, and $L \in \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ is the observer gain selected such that the eigenvalues of the matrix $A-LC$ are strictly within the unit circle. To stabilize the closed-loop process, a linear control law utilizing the state estimate is synthesized:

$$u(t) = -K\hat{x}(t) \tag{4.4}$$

where $K \in \mathbb{R}^{n_u} \times \mathbb{R}^{n_x}$ is the controller gain selected such that the eigenvalues of the matrix $A - BK$ are strictly within the unit circle. The estimation error is defined as the difference between the process state and the state estimate, i.e., $e := x - \hat{x}$, with dynamics given by:

$$e(t+1) = L(I - \Lambda)Cx(t) + (A - LC)e(t) + Gw(t) - L\Lambda v(t) \tag{4.5}$$

To analyze the stability of the closed-loop process under an attack, an augmented state vector is defined as a concatenation of the state and the error vectors $\xi := [x^T \; e^T]^T$. The augmented state dynamics is described by:

$$\xi(t+1) = \underbrace{\begin{bmatrix} (A - BK) & BK \\ L(I - \Lambda)C & (A - LC) \end{bmatrix}}_{=:A_\xi(\Lambda,K,L)} \xi(t) + \underbrace{\begin{bmatrix} G & 0 \\ G & -L\Lambda \end{bmatrix}}_{=:B_\xi(\Lambda,L)} d(t) \tag{4.6}$$

where $d(t) := [w^T(t) \; v^T(t)]^T \in F$ and $F := \{[w^T \; v^T]^T \mid w \in W, v \in V\}$. Due to persistent bounded disturbances acting upon the process, the closed-loop process is continuously perturbed, and the augmented state never converges to the origin. Instead, the augmented state of the closed-loop process is ultimately bounded within a small neighborhood of the origin when $\max_i|\lambda_i(A_\xi(\Lambda, K, L))| < 1$. This neighborhood of the origin is the minimum invariant set of the process and can be expressed as the infinite Minkowski sum [77]:

$$D_\xi(\Lambda, K, L) = \bigoplus_{i=0}^{\infty} A_\xi(\Lambda, K, L)^i B_\xi(\Lambda, L) F \tag{4.7}$$

From Eq. 4.6, the matrices $A_\xi(\Lambda, K, L)$ and $B_\xi(\Lambda, L)$ are dependent on the attack magnitude ($\Lambda$) and the controller-observer parameters $(K, L)$. Consequently, the minimum invariant set in Eq. 4.7 is dependent on the attack magnitude, the controller-observer parameters, and the disturbance set. For simplicity of presentation, the closed-loop process in Eq. 4.1 operated with the control input in Eq. 4.4 computed based on the state estimates and with the controller gain $K$ and the observer gain $L$, is referred to as the closed-loop process with $(K, L)$.

## 4.2 Class of Attack Detection Schemes

The detectability of an attack on the closed-loop process with $(K, L)$ may be defined with respect to the detection scheme monitoring the process. A general class of detection schemes monitoring the process utilizing a generalized monitoring variable $\eta \in \mathbb{R}^{n_\eta}$ is considered. The generalized monitoring variable may be expressed as a weighted combination of the measured output and the estimate of the measured output:

$$\eta(t) = H_y y(t) + H_{\hat{y}} \hat{y}(t) \tag{4.8}$$

where $H_y$ and $H_{\hat{y}}$ are matrices of appropriate dimensions. From Eq. 4.2 and Eq. 4.3b, the measured output and its estimate may also be expressed in terms of the augmented state $\xi(t)$ and the process disturbance $d(t)$ as:

$$y(t) = \underbrace{\begin{bmatrix} \Lambda C & 0 \end{bmatrix}}_{=:A_y(\Lambda)} \xi(t) + \underbrace{\begin{bmatrix} 0 & \Lambda \end{bmatrix}}_{=:B_y(\Lambda)} d(t) \tag{4.9a}$$

$$\hat{y}(t) = \underbrace{\begin{bmatrix} C & -C \end{bmatrix}}_{=:A_{\hat{y}}} \xi(t) + \underbrace{\begin{bmatrix} 0 & 0 \end{bmatrix}}_{=:B_{\hat{y}}} d(t) \tag{4.9b}$$

Thus, Eq. 4.8 may be re-written as:

$$\eta(t) = A_\eta(\Lambda)\xi(t) + B_\eta(\Lambda)d(t) \tag{4.10}$$

where $A_\eta(\Lambda) = H_y A_y(\Lambda) + H_{\hat{y}} A_{\hat{y}}$ and $B_\eta(\Lambda) = H_y B_y(\Lambda) + H_{\hat{y}} B_{\hat{y}}$.

When the closed-loop process with $(K, L)$ is stable in the sense that all eigenvalues of the matrix $A_\xi(\Lambda, K, L)$ are strictly within the unit circle, the augmented state of the process is ultimately bounded within its minimum invariant set $(D_\xi(\Lambda, K, L))$. Furthermore, because the closed-loop process is subjected to bounded disturbances, the generalized monitoring variable is also bounded within a terminal set, denoted by $D_\eta(\Lambda, K, L)$. From Eq. 4.10, the terminal set of the generalized monitoring variable may be computed by:

$$D_\eta(\Lambda, K, L) = A_\eta(\Lambda) D_\xi(\Lambda, K, L) \oplus B_\eta(\Lambda) F \tag{4.11}$$

The generalized monitoring variable is bounded within its attack-free terminal set, i.e., $\eta(t) \in D_\eta(I, K, L)$ for all time $t \geq 0$ if $\xi(0) \in D_\xi(I, K, L)$ because $D_\xi(I, K, L)$ is an invariant set, i.e., $\xi(t) \in D_\xi(I, K, L)$ for all time $t \geq 0$ if $\xi(0) \in D_\xi(I, K, L)$. The class of detection schemes considered in this work monitor the process for attacks by verifying the containment of the generalized monitoring variable within its attack-free terminal set:

$$h(\eta(t)) = \begin{cases} 0, & \eta(t) \in D_\eta(I, K, L) \\ 1, & \text{Otherwise} \end{cases} \tag{4.12}$$

where the mapping $h : \mathbb{R}^{n_\eta} \to \{0, 1\}$ returns the output of the detection scheme, with an output value of 1 being indicative of an attack detection, and an output value of 0 being indicative of a lack of attack detection. The approach adopted herein for tuning the general class of detection schemes accounts for all possible values of process disturbances and measurement noise acting on the process. As a result, the tuning approach adopted ensures a zero false alarm rate in the attack-free process.

One example of a measured variable that fits the model for the generalized detection scheme in Eq. 4.8 is the residual, which measures the deviation of the measured output from its estimate:

$$r(t) := y(t) - \hat{y}(t) = \underbrace{\begin{bmatrix} (\Lambda - I)C & C \end{bmatrix}}_{=:A_r(\Lambda)} \xi(t) + \underbrace{\begin{bmatrix} 0 & \Lambda \end{bmatrix}}_{=:B_r(\Lambda)} d(t) \tag{4.13}$$

Residual-based detection schemes monitor a process utilizing the residual. They are typically used for fault detection [85–87] and have also been extensively explored for attack detection [22, 23, 58, 59, 64, 89]. From Eq. 4.13, the residual fits within the model for the generalized monitoring variable in Eq. 4.8, with $H_y = I$, $H_{\hat{y}} = -I$.

In Ref. 89, an approach to classify attacks based on their detectability with respect to a residual-based detection scheme of the form in Eq. 4.12 was presented. The detectability-based classification of attacks may be extended to a general class of detection schemes of the form in Eq. 4.12 utilizing a monitoring variable of the form in Eq. 4.10. With respect to a class of detection schemes in Eq. 4.12 utilizing a generalized monitoring variable of the form in Eq. 4.10, an attack is said to be detected at time $t_d$ if $\eta(t_d) \notin D_\eta(I, K, L)$ with the output of the detection scheme $h(\eta(t_d)) = 1$. An attack is defined as a detectable attack with respect to the detection scheme in Eq. 4.12 if the attack is detected in finite time (for all $\xi(0) \in \mathbb{R}^{2n_x}$ and $d(t) \in F$ for $t \geq 0$). An attack is defined as an undetectable attack with respect to the detection scheme in Eq. 4.12 if the generalized monitoring variable for the attacked closed-loop process satisfies $\eta(t) \in D_\eta(I, K, L)$ for all $t \geq 0$ for all $\xi(0) \in D_\xi(\Lambda, K, L)$ and $d(t) \in F$ for all $t \geq 0$. Finally, an attack is defined as potentially detectable with respect to the detection scheme in Eq. 4.12 if the attack is neither detectable nor undetectable.

Typically, attack detection schemes using the residual as a monitoring variable have been considered in the literature [22, 23, 58, 59, 64, 89]. However, monitoring both the measured output and the residual may be beneficial for the detection of attacks. For example, an attack ($\Lambda \neq I$) may be undetectable with respect to a residual-based detection scheme with $D_r(\Lambda, K, L) \subseteq D_r(I, K, L)$. However, the attack may be potentially detectable with respect to an output-based detection scheme $D_y(\Lambda, K, L) \nsubseteq D_y(I, K, L)$. As a result, the attack may not be detected by the residual-based detection scheme, but the output-based detection scheme may detect the attack. Similarly, attacks that are undetectable with respect to an output-based detection scheme may be detected by a residual-based detection scheme. In the present work, a detection scheme of the form of Eq. 4.12 monitoring the process using an output and residual-based monitoring variable defined as a concatenation of the measured output and the residual ($\chi := \begin{bmatrix} y^T & r^T \end{bmatrix}^T$) is considered. The monitoring variable $\chi(t) \in \mathbb{R}^{2n_x}$ fits the model for the generalized monitoring variable in Eq. 4.8 with $H_y = \begin{bmatrix} I \\ I \end{bmatrix}$ and $H_{\hat{y}} = \begin{bmatrix} 0 \\ -I \end{bmatrix}$. Therefore, the detectability-based classification of attacks is valid for an output and residual-based detection scheme of the form:

$$h(\chi(t)) = \begin{cases} 0, & \chi(t) \in D_\chi(I, K, L) \\ 1, & \text{Otherwise} \end{cases} \tag{4.14}$$

where $D_\chi(I, K, L)$ is the terminal set of the output and residual-based monitoring variable $\chi$ for the attack-free process. $D_\chi(I, K, L)$ may be computed using Eq. 4.11.

## 4.3 Active Detection Method

In this section, the proposed switching-enabled active detection method for false alarm minimization is presented. A rigorous analysis is employed to develop a switching condition to minimize false alarms.

### 4.3.1 Controller Switching for Active Detection

From the detectability-based classification of attacks, controller-observer parameters, selected to meet standard design criteria, may mask some sensor-controller link multiplicative attacks in the sense that attacks are undetectable with respect to the detection scheme in Eq. 4.14. The controller-observer parameters selected based on standard design criteria are called the nominal controller-observer parameters and are denoted by $(K^*, L^*)$. Other controller-observer parameters may not mask the attacks, making the attacks potentially detectable or detectable with respect to the detection scheme. For the attack-free process, using other controller-observer parameters may lead to performance degradation relative to the closed-loop performance achieved under the nominal controller-observer parameters. Occasional switching between the nominal controller-observer parameters and other controller-observer parameters may be a way to balance the potential trade-off between closed-loop performance and attack detectability. Controller-observer parameter switching is an active detection method because switching probes for multiplicative attacks. The second set of controller-observer parameters is selected to be "sensitive" to attacks over a range of magnitudes, meaning that a range of multiplicative attacks destabilizes the closed-loop process, rendering the attacks detectable. These controller-observer parameters are called attack-sensitive parameters and are denoted by $(K_\Lambda, L_\Lambda)$. The dwell-time under the attack-sensitive controller-observer parameters manages the trade-off between attack detection and performance degradation and is denoted by $T_c$.

The terminal set of the monitoring variable under the attack-sensitive controller-observer parameters is different from the set under the nominal controller-observer parameters. To account for this difference in terminal sets, a time-dependent tuning strategy is used for the detection scheme in Eq. 4.14:

$$h(\chi(t)) = \begin{cases} 0, & \chi(t) \in D_\chi(I, K(t), L(t)) \\ 1, & \text{Otherwise} \end{cases} \tag{4.15}$$

where $(K(t), L(t)) = (K_\Lambda, L_\Lambda)$ for $t \in (t_s, t_s + T_c]$, $(K(t), L(t)) = (K^*, L^*)$ otherwise, $t_s$ denotes the time instance that the control system switches from the nominal controller-observer parameters to the attack-sensitive controller-observer parameters, and $t_s^* = t_s + T_c$ denotes the time instance that the control system switches from the attack-sensitive controller-observer parameters back to the nominal controller-observer parameters.

**Remark 4.3.1.** *The attack-sensitive controller-observer parameters are selected such that undetectable multiplicative sensor-controller link attacks on the process under the nominal controller-observer parameters are rendered detectable under the attack-sensitive parameters. However, finding one pair of controller-observer parameters that renders all attacks detectable may not be possible. Additionally, some attacks that are undetectable under the nominal controller-observer parameters may result in minimal performance deterioration when compared to that under attack-free conditions. Therefore, performance-based selection criteria could be employed to determine the attack-sensitive controller-observer parameters. Multiple attack-sensitive controller-observer parameter pairs may be selected and used to cover a wide range of attacks.*

**Remark 4.3.2.** *For the practical selection of the attack-sensitive controller-observer pa-*

*rameters, a finite set of attacks should be considered. For example, a subclass of multiplicative sensor-controller link attacks may be considered where the attack magnitude may be modeled by a diagonal matrix ($\Lambda = diag(\alpha_1, \dots, \alpha_{n_x})$) and $\alpha_i$ represents the magnitude of the multiplicative attack targeting the $i^{th}$ sensor-controller link. For this subclass of attacks, a finite set of attacks generated by considering a range of values for $\alpha_i$ for each $i$ and $\alpha_j = 1$ for $j \neq i$. Knowledge of prior attacks or attacks that are critical to detect may also be employed for generating the set of attacks for the attack-sensitive parameter selection. The attack detectability under the nominal controller-observer parameters may be verified for each attack to generate a set of undetectable attacks. The resulting set of attacks may be further refined by considering a performance-based criterion. Specifically, the set of attacks may be refined to consider attacks that are such that the radius of the minimum bounding ball of the terminal set of states of the attacked process is greater than (or much greater than) the radius of the minimum bounding ball of the terminal set of states for the attack-free process, i.e., $R(D_x(\Lambda, K^*, L^*)) > R(D_x(I, K^*, L^*))$ where $R(D_x(\Lambda, K^*, L^*)) \coloneqq \max_{x' \in D_x(\Lambda, K^*, L^*)} \|x'\|$ and $D_x(\Lambda, K^*, L^*) = \begin{bmatrix} I & 0 \end{bmatrix} D_\xi(\Lambda, K^*, L^*)$.*

### 4.3.2 Confidence Region-Based Switching Condition to Eliminate False Alarms

Under the proposed active detection method the control system switches between two modes of operation: the nominal mode under which the process is operated with nominal controller-observer parameters, and the attack-sensitive mode under which the process is operated with the attack-sensitive controller-observer parameters. In the attack-free process under the nominal mode, no false alarms are expected due to the tuning approach adopted for the detection scheme in Eq. 4.14. However, switching the control system operating mode on the attack-free process may cause the augmented state to evolve outside the minimum invariant set under the controller-observer parameters for the new mode, potentially resulting in false alarms. For example, consider that the control system switches from the nominal to the attack-sensitive mode at time $t_s$. If $\xi(t_s) \notin D_\xi(I, K_\Lambda, L_\Lambda)$ (this occurs when $\xi(t_s) \in D_\xi(I, K^*, L^*) \setminus D_\xi(I, K_\Lambda, L_\Lambda)$), the augmented state will evolve outside $D_\xi(I, K_\Lambda, L_\Lambda)$ for some time as it converges to $D_\xi(I, K_\Lambda, L_\Lambda)$. The variable $\chi$ during this period may be outside its terminal set ($\chi(t) \notin D_\chi(I, K_\Lambda, L_\Lambda)$ for some $t \geq t_s$), generating false alarms.

The detection objective of the active detection method is to determine if the process is under an attack, or if it is attack-free. False alarms complicate this determination. False alarms may be avoided if the control system switches when the augmented state is in the minimum invariant set under the controller-observer parameters for the new mode. However, the augmented state is not measured directly, so the exact value of the augmented state is unknown. Instead, a region in the augmented state-space containing the augmented state of the attack-free closed-loop process may be constructed to address this issue. This region is time-dependent and can be computed online from the disturbance set ($F$), the measured output, and the residual. The region is called the confidence region and is denoted by $\Xi(K, L, t)$, highlighting the time and the controller-observer parameter pair $(K, L)$ dependence. Based on its definition, the vector $\chi(t)$ may be expressed in terms of the augmented state and disturbance, as:

$$\chi(t) = \underbrace{\begin{bmatrix} C & 0 \\ 0 & C \end{bmatrix}}_{=:\tilde{C}} \xi(t) + \underbrace{\begin{bmatrix} 0 & I \\ 0 & I \end{bmatrix}}_{=:\tilde{D}} d(t) \tag{4.16}$$

From Eq. 4.16, the confidence region can be computed by:

$$\Xi(K, L, t) = \tilde{C}^{-1} \left( \{\chi(t)\} \ominus \tilde{D}F \right) \tag{4.17}$$

The matrix $\tilde{C}$ is invertible because $C$ is invertible.

A few properties are established to develop a switching condition that, when satisfied, leads to zero false alarms from control system switching. First, the relationship between the confidence region, the augmented state, and the minimum invariant set for the attack-free process is established.

**Proposition 6.** *Consider the attack-free closed-loop process with $(K, L)$. If the matrix $C$ is invertible and $\xi(0) \in D_\xi(I, K, L)$, then the confidence region $\Xi(K, L, t)$ contains the augmented state, i.e., $\xi(t) \in \Xi(K, L, t)$. Furthermore, the confidence region has a non-empty intersection with the minimum invariant set, i.e., $\Xi(K, L, t) \cap D_\xi(I, K, L) \neq \emptyset$.*

From Eq. 4.16, the confidence region is computed under the assumption that the process is attack-free, and therefore, the augmented state will be contained in the confidence region of the attack-free process. If the process is under a cyberattack, the confidence region does not give any information about the value of the augmented state. However, if the confidence region does not intersect the attack-free minimum invariant set, the process cannot be attack-free, because of an inconsistency between the computation of the confidence region for the attack-free process and the expected evolution of the attack-free process state within the minimum invariant set. In this regard, the confidence region may be another mechanism for detecting attacks. In particular, an attack can be declared if the confidence region and the minimum invariant set do not intersect. This is formally stated in the following proposition.

**Proposition 7.** *Consider the closed-loop process with $(K, L)$. Let the matrix $C$ be invertible and $\xi(0) \in D_{\xi}(I, K, L)$. If the confidence region does not intersect with the minimum invariant set of the attack-free closed-loop process, i.e., $\Xi(K, L, t) \cap D_{\xi}(I, K, L) = \emptyset$, then the process is not attack-free.*

Proposition 7 provides a confidence region-based condition that may be verified to monitor a process for attacks. However, the motivation behind constructing the confidence regions is to ensure zero false alarms from a switch between any two controller-observer parameter pairs $(K_1, L_1)$ and $(K_2, L_2)$. To ensure zero false alarms, the augmented state at the switching instance of the attack-free process must be within the attack-free minimum invariant sets under both controller-observer parameters. Based on this, the following theorem leverages the result of the Proposition 6 to establish a condition that, if satisfied at the time instance when the controller-observer parameters switch between $(K_1, L_1)$ to $(K_2, L_2)$, guarantees that zero false alarms are generated in the detection scheme in Eq. 4.15. This further implies that any alarms generated are the result of an attack.

**Theorem 2.** *Consider the closed-loop process with $(K_1, L_1)$. Let the matrix $C$ be invertible and $\xi(0) \in D_{\xi}(I, K_1, L_1)$. Assume that a controller-observer parameter switch from $(K_1, L_1)$ to $(K_2, L_2)$ occurs at $t_s$. If the closed-loop process is attack-free and the confidence region satisfies $\Xi(K_1, L_1, t_s) \cap D_{\xi}(I, K_1, L_1) \subseteq D_{\xi}(I, K_2, L_2)$, then no alarms are* *generated by the detection scheme of the form in Eq. 4.15. Furthermore, if there is an alarm generated by the detection scheme at some time $t_d$, then the closed-loop process is not attack-free.*

These results provide insight into how to design a confidence region-based switching condition. To implement the active detection method without false alarms, a switching condition can be imposed at each switch. When the control system switches from the nominal mode to the attack-sensitive mode at $t_s$, the confidence region should satisfy

$$\Xi(K^*, L^*, t_s) \cap D_\xi(I, K^*, L^*) \subseteq D_\xi(I, K_\Lambda, L_\Lambda) \tag{4.18}$$

When the control system switches from the attack-sensitive mode back to the nominal mode at $t_s^* = t_s + T_c$, the confidence region should satisfy

$$\Xi(K_\Lambda, L_\Lambda, t_s^*) \cap D_\xi(I, K_\Lambda, L_\Lambda) \subseteq D_\xi(I, K^*, L^*) \tag{4.19}$$

### 4.3.3 Minimizing False Alarms

In prior work [89], an active detection method utilizing a time-triggered control system switching approach was presented. Under a time-triggered switching approach, the switching instance $t_s$ and the dwell-time $T_c$ are predetermined. However, process disturbances and measurement noise affect the evolution of the augmented state. At $t_s$ and $t_s^*$, the desired switching conditions in Eq. 4.18 and Eq. 4.19, respectively, may not be satisfied. Also, the existence of $t_s$ and $t_s^*$ when Eq. 4.18 and Eq. 4.19 are satisfied cannot be guaranteed in general. To minimize false alarms, a state-dependent control system switching approach is utilized in the present work. Specifically, an interval of switching times is defined, over which the desired switching condition is verified. If the switching condition is satisfied, the control system switch occurs. In this sense, the switching times may be considered to be state-dependent. If the condition is not satisfied, the operator may choose to force the switch to occur or reschedule it.

For the switch from the nominal mode to the attack-sensitive mode, an interval is defined and is denoted by $[t_i, t_f]$ where $t_i \geq 0$ and $t_f > t_i$ are lower and upper bounds of the interval, respectively. Beginning at $t_i$, the switching condition in Eq. 4.18 is verified at every time step. If the condition is satisfied at $t_s \in [t_i, t_f]$, the control system switches from the nominal mode to the attack-sensitive mode. If the condition is never satisfied over the interval $[t_i, t_f]$, the process operator has a few options. The operator may choose to force the switch to the attack-sensitive mode to occur at $t_f$ or re-schedule the switch to another time. For scheduling the switch to attack-sensitive mode, several factors could be considered. For example, the interval may be chosen as the time interval when the performance degradation resulting from operating with the attack-sensitive mode is acceptable. If operational considerations allow for an unbounded implementation interval, i.e., $t_f \to \infty$, the closed-loop process with $(K^*, L^*)$ may be monitored for an appropriate switching instance over an extended period.

A similar range of switching instances is defined for switching back to the nominal mode. Denoting the minimum and maximum dwell-time under the attack-sensitive mode by $T_c^{min}$ and $T_c^{max}$, respectively, the range of switching instances is given by $[t_s + T_c^{min}, t_s + T_c^{max}]$, i.e., $T_c \in [T_c^{min}, T_c^{max}]$ and $t_s^* \in [t_s + T_c^{min}, t_s + T_c^{max}]$. Starting at $t_s + T_c^{min}$, the condition in Eq. 4.19 is checked. If satisfied at $t_s^*$, the switch is performed. If the condition is never satisfied over the interval, the control system switches back to the nominal mode at $t_s + T_c^{max}$, to minimize the performance degradation. However, false alarms are possible in this case. For the selection of the switching interval, operating the process with the attack-sensitive mode for as long as possible may be desirable from an attack detection perspective. However, limiting the dwell-time under the attack-sensitive mode may be desirable to limit performance degradation. Thus, $T_c^{min}$ and $T_c^{max}$ manage the trade-off between attack detection and performance degradation. For example, the minimum dwell-time $T_c^{min}$ may be chosen as the period for which most attacks on the process in the attack-sensitive mode are detected, as demonstrated in the illustrative case study section. Similarly, the maximum dwell-time specifies a limit to the operation in attack-sensitive mode. To this end, $T_c^{max}$ may be selected as the time of operation in attack-sensitive mode while maintaining process states within a safe set.

Under the proposed active detection method, an operator may choose to force a control system switch at a time when the zero false alarm condition in Eq. C.2 is not satisfied. In the event of a forced control system switch on the attack-free process, false alarms may be generated for a few time steps until the augmented state converges to the minimum invariant set under the updated controller-observer parameters. Therefore, to minimize false alarms, a modification to the detection scheme in Eq. 4.14 may be considered. Under the modified detection scheme, alarms generated after a forced control system switch may be suppressed for a few time steps. This suppression of alarms is in-line with the standard industry practice of adding a delay timer to the alarm logic of a controller [88]. After the period for suppression of alarms elapses, any alarm generated in the detection scheme in Eq. 4.14 may be considered to be indicative of the detection of an attack.

As part of the proposed active detection method, in addition to the detection scheme in Eq. 4.15, the confidence regions are used to monitor the process for an attack (leveraging the result of Proposition 7):

$$z(t) = \begin{cases} 0, \ \Xi(K(t), L(t), t) \cap D_{\xi}(I, K(t), L(t)) \neq \emptyset \\ 1, \ \text{Otherwise} \end{cases} \tag{4.20}$$

where $z(t) \in \{0, 1\}$ is the output of the detection scheme, with an output of 1 being indicative of attack detection, and an output of 0 indicating a lack of attack detection. Algorithm 1 covers the monitoring logic, control system switching logic, and control action computation over a single cycle switching into and out of the attack-sensitive mode under the proposed active detection method.

The algorithm inputs are the time interval for switching into the attack-sensitive mode ($[t_i, t_f]$), the dwell-time range under the attack-sensitive mode ($[T_c^{min}, T_c^{max}]$), the alarm suppression time after a forced switch into the attack-sensitive mode ($\Delta_1$), the alarm suppression time after a forced switch back from attack-sensitive mode ($\Delta_2$), and the nominal controller-observer parameters ($K^*, L^*$), and the attack-sensitive controller-observer parameters ($K_\Lambda, L_\Lambda$). To perform some computations in the algorithm, additional parameters are needed ($D_\xi(I, K^*, L^*)$, $D_\xi(I, K_\Lambda, L_\Lambda)$, $D_\chi(I, K^*, L^*)$, and $D_\chi(I, K_\Lambda, L_\Lambda)$). These parameters have been omitted for simplicity of presentation. Without loss of generality, the algorithm is activated at time $t_i$. The algorithm terminates when the control system

**Algorithm 1:** Algorithm for the active detection method

```
Inputs: $t_i < t_f$, $\Delta_1 < T_c^{min} \leq T_c^{max}$, $\Delta_2$, $(K^*, L^*)$, $(K_\Lambda, L_\Lambda)$
Initialization: $t = t_i$, $t_s = \infty$, $t_s^* = \infty$, $t_d = \infty$, $\Delta(t) = 0$, $(K(t), L(t)) = (K^*, L^*)$
Outputs: $t_d$, $t_s$, $t_s^*$
while $t \leq t_s^* + \Delta_2$ do
    Receive the measured output $y(t)$ communicated over the sensor-controller link
    Compute the residual $r(t)$ and the confidence region $\Xi(K(t), L(t), t)$ from
      $\chi(t) = \left[ y^T(t) \; r^T(t) \right]^T$
    Monitoring logic
    if $h(\chi(t)) = 1$ or $z(t) = 1$ then
        if $\Delta(t) = 0$ then
            An attack is detected. Set $t = t_d$
            Activate attack identification and mitigation strategies
        else
            Suppress alarms. Set $h(\chi(t)) = 0$ and $z(t) = 0$
        end
    end
    Switching logic
    if $t_s = \infty$ and $t \in [t_i, t_f]$ then
        if Eq. 4.18 is satisfied then
            Switch to attack-sensitive mode. Set $t_s = t$ and $(K(t), L(t)) = (K_\Lambda, L_\Lambda)$
        else if $t = t_f$ then
            Switch to attack-sensitive mode. Set $t_s = t$, $(K(t), L(t)) = (K_\Lambda, L_\Lambda)$, and
              $\Delta(t) = \Delta_1$
        end
```

**Algorithm 1:** Algorithm for the active detection method

```
    else if $t_s \neq \infty$, $t_s^* = \infty$, and $t \in [t_s + T_c^{min}, t_s + T_c^{max}]$ then
        if Eq. 4.19 is satisfied then
            Switch to nominal mode. Set $t_s^* = t$ and $(K(t), L(t)) = (K^*, L^*)$
        else if $t = t_s + T_c^{max}$ then
            Switch to nominal mode. Set $t_s^* = t$, $(K(t), L(t)) = (K^*, L^*)$, and
              $\Delta(t) = \Delta_2$
    Compute the control action $u(t)$
    Communicate the computed control action to the actuators
    Set $t \leftarrow t+1$, $(K(t+1), L(t+1)) = (K(t), L(t))$, and $\Delta(t+1) = \max\{\Delta(t) - 1, 0\}$
```

switches back to the nominal mode or when an attack is detected. If an attack is detected, attack identification and mitigation strategies are activated, albeit a discussion of these strategies is beyond the scope of the current work. The variable $\Delta(t)$ tracks the number of time steps from the time step $t$ that any alarms should be suppressed. To ensure that the switch back into the nominal mode does not occur during the alarm suppression period, the alarm suppression period after a forced switch into attack-sensitive mode is chosen to be less than the minimum dwell-time, i.e., $\Delta_1 < T_c^{min}$. The algorithm outputs are the detection time and the switching instances.

When the algorithm is not active, the process is assumed to be operated and monitored under the nominal mode. The algorithm may be periodically activated, enabling routine cyberattack probing. Additionally, the algorithm may be activated multiple times using different attack-sensitive controller-observer parameters to probe for different attacks. No attacks are assumed to be detected before activating the algorithm because switching into attack-sensitive mode is not needed if an attack is detected before the algorithm is activated.

**Remark 4.3.3.** *After a forced control system switch between any two parameter modes,* *i.e., from controller-observer parameters $(K_1, L_1)$ to controller-observer parameters $(K_2, L_2)$ on the attack-free process, a conservative estimate of the alarm suppression time ($\Delta'$) for the detection scheme in Eq. 4.14 may be estimated as the time taken by any realization of the augmented state within the minimum invariant set of the process under $(K_1, L_1)$ to converge to the minimum invariant set of the process under $(K_2, L_2)$.*

**Remark 4.3.4.** *The set of attack magnitudes, i.e., the set of values of $\Lambda \neq I$, that may be detected under a given control mode (i.e., attack-sensitive mode or the nominal mode) is the set of potentially detectable or detectable attacks. Since the process model and admissible set of process disturbances and measurement noise are fixed, this set is only dependent on the controller-observer parameters of the active control mode. The set of attack magnitudes that will be detected under a given control mode depends on the controller-observer parameters and other factors, including the dwell-time under the active mode and the realizations of the process disturbance and measurement noise. The set of attacks that may be detected can be numerically approximated by checking the detectability of attacks within a finite set of values, although the accuracy of this approximation may be limited by the number of attack magnitudes considered. However, an explicit characterization of the set of attacks that will be detected is an open problem.*

## 4.4 Illustrative Case Studies

In this section, two illustrative processes are considered to demonstrate the application of the active detection method. All polytope computations are performed using the Multi-Parametric Toolbox (MPT 3.0) [82].

### 4.4.1 Application to a Scalar Process

A scalar process consisting of a single state ($x(t) \in \mathbb{R}$), and a single measured output ($y(t) \in \mathbb{R}$) is considered:

$$x(t+1) = x(t) + u(t) + w(t)$$

$$y(t) = \Lambda(x(t) + v(t))$$

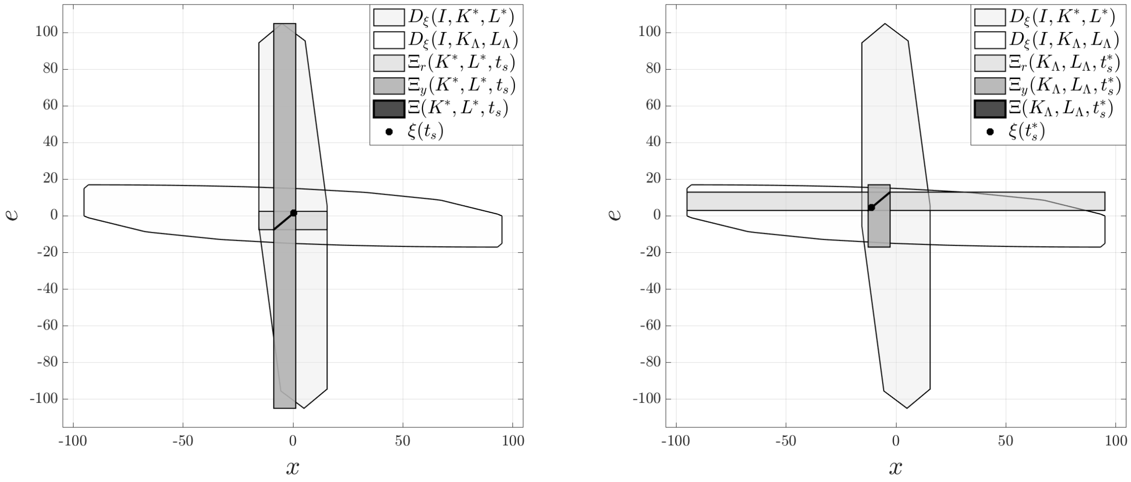
where $u(t) \in \mathbb{R}$ is the manipulated input, $\Lambda \neq 1$ is the magnitude of multiplicative sensor-controller attack, $v(t) \in V := \{v' \mid v' \in [-5, 5]\}$ represents the vector of bounded measurement noise corrupting the measurements of the state, and $w(t) \in W := \{w' \mid w' \in [-1, 1]\}$ represents the vector of bounded process disturbances. A Luenberger observer of the form in Eq. 4.3a is synthesized to generate estimates of states $\hat{x}(t) \in \mathbb{R}$. To stabilize the process at the origin, which is the desired operating steady-state, a linear feedback law of the form Eq. 4.4 is used to compute the control input from the estimates of state. To analyze the stability of the closed-loop process, an augmented state vector $\xi := \begin{bmatrix} x \; e \end{bmatrix}^T$ is defined. The closed-loop process is expressed in the form of Eq. 4.6 with

$$A_\xi(\Lambda, K, L) = \begin{bmatrix} (1-K) & K \\ L(1-\Lambda) & 1-L \end{bmatrix}, \; B_\xi(\Lambda, K, L) = \begin{bmatrix} 1 & 0 \\ 1 & -L\Lambda \end{bmatrix}$$

where $K$ is the controller gain and $L$ is the observer gain.

The nominal controller-observer parameters for the process are chosen as $K^* = 0.1$ and $L^* = 1.9$ to stabilize the attack-free closed-loop process. To detect attacks with magnitudes in the range $\Lambda \in [1.3, 4]$, the attack-sensitive controller-observer parameters for the process are chosen with $K_\Lambda = 1.7$ and $L_\Lambda = 1.5$. The range of attacks that destabilize the closed-loop process under attack-sensitive controller-observer parameters is numerically verified by checking if the value of $\max_i |\lambda_i(A_\xi(\Lambda, K_\Lambda, L_\Lambda))| > 1$ for all $\Lambda \in [1.3, 4]$, by starting at an attack magnitude equal to the lower bound of the range ($\Lambda = 1.3$), and incrementing the magnitudes by 0.01 until the upper bound of the range is reached ($\Lambda = 4$). A similar analysis performed for nominal controller-observer parameters reveals that they are not sensitive to attacks in the interval $[1.3, 4]$. For the attack-free process under the nominal and the attack-sensitive mode, the radii of the minimum bounding balls containing the terminal set of states are computed as $R(D_x(I, K^*, L^*)) = 15.5263$ and $R(D_x(I, K_\Lambda, L_\Lambda)) = 95$ where $D_x(I, K, L)$ denotes the terminal set of states for the attack-free closed-loop process with parameters $(K, L)$. Defining closed-loop performance with the radius of the minimum bounding ball containing the terminal set, the closed-loop performance under the nominal parameters is better than that under the attack-sensitive parameters.

To monitor the process using a detection scheme of the form of Eq. 4.15, invariant outer approximations of the minimum invariant sets of the attack-free process under the nominal and the attack-sensitive controller-observer parameters are computed as $D_\xi(I, K^*, L^*)$ and $D_\xi(I, K_\Lambda, L_\Lambda)$ using the method described in Ref. 79. The error bound used in computing the numerical approximations is $\epsilon = 5 \times 10^{-5}$. Numerical approximations of the sets $D_\chi(I, K^*, L^*)$ and $D_\chi(I, K_\Lambda, L_\Lambda)$ are computed from Eq. 4.11 and shown in Fig. 4.2. The confidence region constructed using the monitored variable $\chi(t)$ is compared with two other methods for computing the confidence region: one using the measured output and one using the residual. From the measured output, a set containing the process state may be computed by: $X_y(K, L, t) = \{y(t)\} \ominus V$ (for a given controller-observer parameter pair $(K, L)$). Since the augmented state of the attack-free process is bounded within its minimum invariant set, the estimation error is bounded within its terminal set, computed by: $D_e(I, K, L) := [0\ 1]D_\xi(I, K, L)$. Therefore, the sets $X_y(K, L, t)$ and $D_e(I, K, L)$ are the regions containing the process state and the estimation error. A confidence region constructed using the output alone is given by: $\Xi_y(K, L, t) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} X_y(K, L, t) \oplus \begin{bmatrix} 0 \\ 1 \end{bmatrix} D_e(I, K, L)$. Similarly, from Eq. 4.13, the residual value for the attack-free process depends on the estimation error and the measurement noise. A set containing the estimation error values may be computed by: $E_r(K, L, t) = \{r(t)\} \ominus V$. The terminal set of states may be computed by: $D_x(I, K, L) = [1\ 0]D_\xi(I, K, L)$. Therefore, the confidence region containing attack-free states constructed from the residual alone may be computed by: $\Xi_r(K, L, t) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} D_x(I, K, L) \oplus \begin{bmatrix} 0 \\ 1 \end{bmatrix} E_r(K, L, t)$. Therefore, the confidence region computed from the output and residual-based monitoring variable may be compared with the confidence region computed from the measured output alone and that computed from the residual alone.



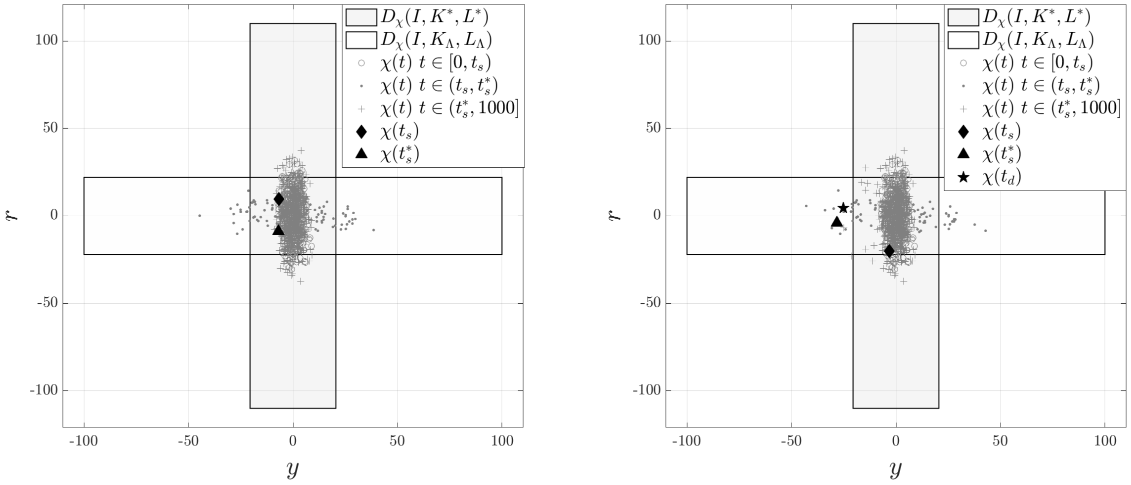
(a) (b)

Fig. 4.1: (a) Confidence regions for the attack-free process under the nominal mode at the time $t_s = 250$. (b) Confidence regions for the attack-free process under the attack-sensitive mode at the time $t_s = 350$.

A simulation of the attack-free scalar process with the proposed active detection method is considered. For the active detection method, the control system switch from the nominal mode to the attack-sensitive mode is scheduled over the interval $[t_i, t_f] = [250, 400]$. Over this interval, the condition in Eq. 4.18 is verified at each time step. The minimum and the maximum dwell-time under attack-sensitive mode are selected to be $T_c^{min} = 100$ and $T_c^{max} = 110$. The process disturbances and measurement noise are modeled as random variables drawn from uniform distributions at each step and bounded between $[-1, 1]$ and $[-5, 5]$, respectively. The total length of the simulation is 1000 time steps, and the initial condition of the process is 0. To implement a switch, a confidence region computed using the monitoring variable $\chi(t)$ is used to check the appropriate switching condition. Over the simulation, the switch from the nominal mode to the attack-sensitive mode occurs at the time step $t_s = 250$ when the condition in Eq. 4.18 is satisfied (Fig. 4.1a). Similarly, the switch back to nominal mode occurs at the time step $t_s^* = t_s + T_c^{min} = 350$ when the condition in Eq. 4.19 is satisfied (Fig. 4.1b). No false alarms are observed due to either switch.

For comparison, the confidence regions are computed from the residual and output at both switching instances and are depicted in Fig. 4.1a and Fig. 4.1b. At both switching instances, the augmented state is contained within the confidence region constructed from the residual and from the output. However, the confidence region computed from the output does not satisfy Eq. C.2 with $\Xi_y(K^*, L^*, t_s) \cap D_\xi(I, K^*, L^*) \nsubseteq D_\xi(I, K_\Lambda, L_\Lambda)$ at the switching instance $t_s = 250$ (Fig. 4.1a). As a result, the switch may have been prevented if the switching condition is verified based on the confidence region computed from the output. Similarly, the confidence region computed from the residual does not satisfy Eq. C.2 with $\Xi_r(K_\Lambda, L_\Lambda, t_s^*) \cap D_\xi(I, K_\Lambda, L_\Lambda) \nsubseteq D_\xi(I, K^*, L^*)$, and may have prevented a switch to the attack-sensitive mode at the time $t_s^*$. Furthermore, when compared to the confidence regions computed from the output and residual-based monitoring variable $\chi(t)$, confidence regions computed from the output or the residual alone are larger regions. Therefore, the confidence region computed from $\chi(t)$ provides a less conservative estimate of the region containing the attack-free augmented state, and is considered in the present work.

Next, the minimization of false alarms in an attack-free process with the proposed active detection method is demonstrated. Two scenarios are considered. The first scenario considers the attack-free process with the proposed active detection method. The second scenario considers the attack-free process with the active detection method, but with a time-triggered control system switching. Each scenario consists of 1000 simulations, where the bounded process disturbances and measurement noise at each time step are drawn from a uniform distribution as described previously. The same realization of the random variables is used in both scenarios to compare across simulations. The initial condition of all simulations is 0, which is contained within the attack-free minimum invariant set under the nominal controller-observer parameters. The total length of each simulation is 1000 time steps.



(a)

(b)

Fig. 4.2: (a) Monitoring variable values for the attack-free process with the proposed active detection method. (b) Monitoring variable values for the attack-free process with a time-triggered control system switching.

In the first scenario, the proposed active detection method is applied to the attack-free process. For the active detection method, the algorithm is implemented with a time interval $[t_i, t_f] = [250, 400]$ for a switch from the nominal mode to attack-sensitive mode, and a dwell-time range $T_c^{min} = 100$ and $T_c^{max} = 110$ for the switch back from the attack-sensitive mode to nominal mode are used. The alarm suppression period after each control system switch is chosen to be 10-time steps, i.e., $\Delta_1 = 10$ and $\Delta_2 = 10$. Over numerous simulations of the attack-free process with a time-triggered switch, the augmented state converges to the minimum invariant set under the new controller-observer parameters within 10-time steps or less. The switch into attack-sensitive mode to probe for attacks is scheduled for $[250, 308]$. The switch back to the nominal mode is implemented over the interval $[350, 412]$. The switch back to nominal mode occurred when the condition in Eq. 4.19 is satisfied in 977 of the 100 simulations. Over 23 of the 1000 simulations, the switch back to nominal mode is forced at the time $t_s^* = t_s + T_c^{max}$ because Eq. 4.19 is not satisfied over the implementation interval. Over the remaining 23 simulations, the augmented state converged to the minimum invariant set under the nominal controller-observer parameters in 10-time steps or less. As a result, no false alarms are observed in the detection scheme in Eq. 4.15.

Fig. 4.2a illustrates the monitoring variable values from one simulation. Over this simulation, the monitoring variable values are contained within the terminal set under nominal controller-observer parameters as indicated by the unfilled circular markers in Fig. 4.2a. When the switch into attack-sensitive mode occurs at time step $t_s = 264$, the monitoring variable value is represented by a diamond marker in Fig. 4.2a. After the switch into the attack-sensitive mode, the monitoring variable values are contained within the corresponding terminal set, as indicated by dot markers in Fig. 4.2a. The switch back to nominal mode occurs at the time $t_s^* = 366$, with a monitoring value represented by a triangle marker in Fig. 4.2a. After the switch, the monitoring variable $\chi(t)$ is contained within its corresponding terminal set, as indicated by the "plus" markers in Fig. 4.2a.

In the second scenario, the attack-free process with an active detection method, but with a time-triggered switching strategy, is considered. The switch into attack-sensitive mode occurs at the time $t_s = 250$, and in the absence of an attack detection, a switch back to the nominal mode occurs at the time $t_s^* = 350$. In 1000 simulations of the process under the time-triggered switching strategy, no false alarms are observed after the switch from the nominal to the attack-sensitive mode. In 204 out of 1000 simulations, false alarms are generated in the detection scheme in Eq. 4.15, after switch back to nominal mode. The monitoring variable values over one simulation are illustrated in Fig 4.2b. As indicated by the unfilled circular markers in Fig. 4.2b, the monitoring variable values are contained within the terminal set under nominal controller-observer parameters until the switch into attack-sensitive mode occurs at the time step $t_s = 264$ (with monitoring variable value represented by a diamond marker in Fig. 4.2b). After the control system switches into attack-sensitive mode, the monitoring variable values are contained within the corresponding terminal set, as indicated by dot markers in Fig. 4.2b. No alarms are observed after switching into attack-sensitive mode at the time step $t_s = 250$. The switch back to the nominal mode occurs at the time $t_s^* = 366$, with a monitoring value represented by a triangle marker in Fig. 4.2b. After the switch, an attack detection (false alarm) is reported by the detection scheme in Eq. 4.14 at the time step $t_d = 351$ (indicated by the filled star marker in Fig. 4.2b). False alarms are observed for up to 2 more time steps, after which the monitoring variable $\chi(t)$ is contained within its corresponding terminal set, as indicated by the "plus" markers in Fig. 4.2b. With the time-triggered switching strategy, false alarms spanning 10 time steps or less are observed in 204 simulations of the process. However, no false alarms are observed over all simulations of the process with the proposed active detection method. Therefore, the proposed active detection method minimizes false alarms from a switch.

A third scenario with the process under an attack of magnitude $\Lambda = 1.3$ with the proposed active detection method is considered to demonstrate enhancement of detection capabilities. The attack is potentially detectable under nominal controller-observer parameters and detectable under attack-sensitive controller-observer parameters. For a basis of comparison, 1000 simulations of the attacked process operated exclusively under the nominal mode are performed. Over 1000 simulations, the attack is not detected by the detection scheme in Eq. 4.14. Next, 1000 simulations of the attacked scalar process with the proposed active detection method are performed. Over 1000 simulations, the attack is detected by the scheme in Eq. 4.15 within a maximum of 47-time steps from the switch into attack-sensitive mode. Thus, the active detection method enhances the detection capabilities of the detection scheme in Eq. 4.14. Additionally, this case study highlights the possible use of monitoring a process using the confidence region-based detection scheme in Eq. 4.20 because the attack is detected by the confidence region-based detection scheme in Eq. 4.20 in all simulations.

### 4.4.2 Application to a Chemical Process

A chemical process consisting of a well-mixed continuously stirred tank reactor (CSTR) where a second-order, single-phase exothermic reaction of the form $A \rightarrow B$ occurs is considered. The tank liquid may be heated or cooled. Applying standard modeling assumptions, the dynamic process model is obtained from the mass and energy balances

Model parameters for the CSTR

| | |
|---|---|
| Density | $\rho_L = 1000\,\mathrm{kg\,m^{-3}}$ |
| Heat capacity | $C_p = 0.231\,\mathrm{kJ\,kg^{-1}\,K^{-1}}$ |
| Flow rate | $F = 5.0\,\mathrm{m^3\,h^{-1}}$ |
| Reactor volume | $V = 1.0\,\mathrm{m^3}$ |
| Heat of reaction | $\Delta H = -1.15 \times 10^4\,\mathrm{kJ\,kmol^{-1}}$ |
| Activation energy | $E = 5.0 \times 10^4\,\mathrm{kJ\,kmol^{-1}}$ |
| Feed temperature | $T_0 = 300.0\,\mathrm{K}$ |
| Pre-exponential factor | $k_0 = 8.46 \times 10^6\,\mathrm{m^3\,kmol^{-1}\,h^{-1}}$ |
| Gas constant | $R = 8.314\,\mathrm{kJ\,kmol^{-1}\,K^{-1}}$ |
| Concentration of reactant $A$ in the feed | $C_{A0} = 4.0\,\mathrm{kmol\,m^{-3}}$ |

around the CSTR liquid hold-up and is given by:

$$\begin{aligned} \frac{dC_A}{dt} &= \frac{F}{V}(C_{A0} + \Delta C_{A0} - C_A) - k_0 e^{\frac{-E}{RT}} C_A^2 \\ \frac{dT}{dt} &= \frac{F}{V}(T_0 + \Delta T_0 - T) - \frac{\Delta H k_0}{\rho C_p} e^{\frac{-E}{RT}} C_A^2 + \frac{Q}{\rho C_p V} \end{aligned} \tag{4.21}$$

where $C_{A0}$ is the inlet concentration of the reactant, $T_0$ is the inlet temperature, $C_A$ is the concentration of the reactant in the reactor, $T$ is the temperature of the reactor, and $Q$ is the heat supplied to/removed from the reactor. The definitions and values of the process parameters in Eq. 4.21 are given in Table 2.1, and are reproduced in this chapter to make it self-contained. The manipulated input is $Q$. The variables $\Delta C_{A0}$ and $\Delta T_0$ represent deviations in the feed conditions from the nominal values, $C_{A0}$ and $T_0$, respectively, and are considered to be bounded process disturbances. The measured variables are the reactant concentration ($C_A$) and temperature ($T$), with additive bounded measurement noise. The output matrix ($C = I$) is invertible.

The control objective of the CSTR process is to operate the process at the steady-state corresponding to $C_{As} = 1.22\,\mathrm{kmol\,m^{-3}}$, $T_s = 438\,\mathrm{K}$, and $Q_s = 0\,\mathrm{kW}$. A state-space model for the process is obtained using the deviation variables $x_1 = C_A - C_{As}$, $x_2 = T - T_s$, and $u = Q - Q_s$, where $x = [x_1\ x_2]^T$ are deviation variables representing the process states, and $u$ is the deviation variable representing the manipulated input. A discrete-time linear model is needed to design the control system and analyze the attack detectability properties. The nonlinear process model is linearized about the steady-state. The resulting continuous-time linear model is discretized assuming zeroth-order hold of the inputs with a sampling period of $\Delta t = 1 \times 10^{-2}\,\mathrm{h}$. The discrete-time state-space matrices are given by:

$$A = \begin{bmatrix} 0.7364 & -0.0041 \\ 10.6953 & 1.1560 \end{bmatrix}, \; B = \begin{bmatrix} -9.0708 \times 10^{-8} \\ 4.6741 \times 10^{-5} \end{bmatrix}, \; G = \begin{bmatrix} 0.0433 & -0.0001 \\ 0.2724 & 0.0540 \end{bmatrix}$$

For the attack detectability analysis, the algorithm presented in Ref. [79] is used to generate outer invariant approximations of the minimum invariant sets for the attack-free closed-loop process. The maximum error of the outer approximations of the minimum invariant sets is set to $5 \times 10^{-5}$. Outer estimates of the terminal sets of the monitoring variable for the attack-free closed-loop process are computed using the estimates of the minimum invariant sets of the process.

The nominal controller-observer parameters $(K^*, L^*)$ are selected to stabilize the closed-loop process using pole placement by placing the poles at $[0.2 \;\; -0.1]$ to determine the controller gain and placing the poles at $[0.2 \;\; 0.3]$ to determine the observer gain. The attack-sensitive controller-observer parameters $(K_\Lambda, L_\Lambda)$ are determined by placing the poles at $[-0.2 \;\; -0.3]$ and $[-0.2 \;\; -0.3]$ to compute the controller and observer gains, respectively. The control system with the attack-sensitive controller-observer parameters is sensitive to attacks in the set: $\{\Lambda \mid \mathrm{diag}(1, \alpha) \mid \alpha \in [0.6, 0.9]\}$. This range of attacks is verified by checking the eigenvalues of the matrix $A_\xi(\Lambda, K_\Lambda, L_\Lambda)$ with $\Lambda = \mathrm{diag}(1, \alpha)$ and varying $\alpha$ starting from $\alpha = 0.6$ and incrementing by 0.01 until a maximum value of $\alpha = 0.9$ is reached. Performing a similar analysis for the nominal controller-observer parameters found that the nominal controller-observer parameters are not sensitive to any attack over the range checked.

The theoretical analysis of this work considered linear systems of the form in Eq. 4.1. The active detection method is applied to a nonlinear process to demonstrate its applicability to a nonlinear process, extending beyond what is considered in the theoretical analysis.

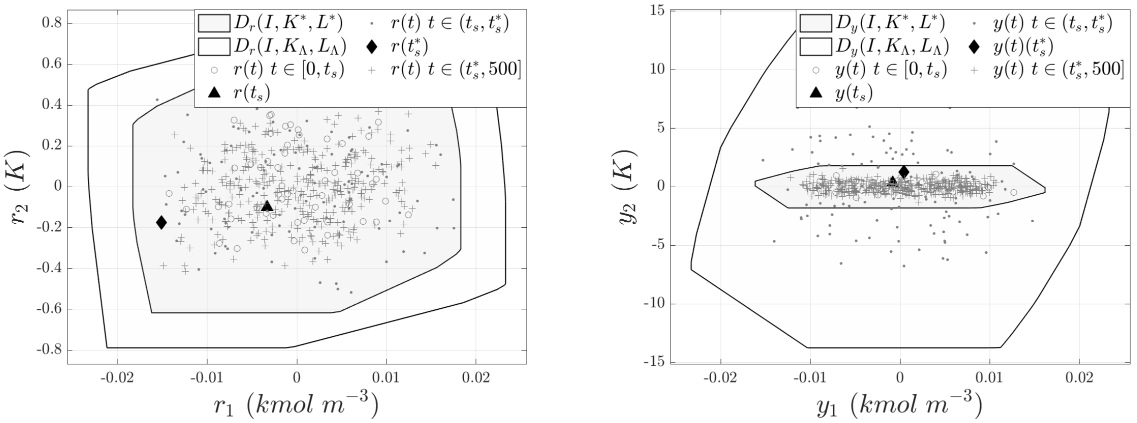
The discrete-time linear control system is applied to the nonlinear process in a sample-and-hold fashion. To integrate the nonlinear ordinary differential equations in Eq. 4.21, the explicit Euler method is used with an integration step size of $1 \times 10^{-4}\,\mathrm{h}$.

Two scenarios are considered. The first scenario considers the attack-free process with the proposed active detection method that minimizes false alarms. The second scenario considers the application of the proposed active detection method to the attacked process to demonstrate the enhancement of detection capabilities of the detection scheme in Eq. 4.15. Each scenario consists of 1000 simulations, where the bounded process disturbances in the feed concentration $\Delta C_{A0}$ and the measurement noise in the concentration sensor are modeled as random numbers drawn from two different uniform distributions on the interval $[-0.01, 0.01]$ $\mathrm{kmol\,m^{-3}}$. Similarly, the bounded process disturbances in the feed temperature $\Delta T_0$ and the measurement noise in the temperature sensor are modeled as random numbers drawn from two different uniform distributions on the interval $[-0.2, 0.2]$ K. The same realization of the random variables is used in each scenario to compare across simulations. The initial condition of all simulations is 0, which is contained within the attack-free minimum invariant set under the nominal controller-observer parameters. The total length of each simulation is 5 h.

In the first scenario, the proposed active detection method is applied to the attack-free CSTR process to demonstrate false alarm minimization. A switch into the attack-sensitive mode to probe for attacks is scheduled for $[t_i, t_f] = [50, 400]$ corresponding to a real-time interval of $[0.5, 4]$ h. The minimum and maximum dwell-time under the attack-sensitive mode are selected to be $T_c^{min} = 100$ (1 h in real-time) and $T_c^{max} = 110$ (1.1 h in real-time). The alarm suppression times are chosen to span 2 time steps from a switch, i.e., $\Delta_1 = 2$ and $\Delta_2 = 2$. This is because the augmented state converges to the minimum invariant set under the new controller-observer parameters in 2-time steps or less after a switch over numerous simulations of the attack-free process with a time-triggered control system switch.

No alarms are raised by the detection scheme in Eq. 4.15 in any of the 1000 simulations of the attack-free process with the proposed active detection method. The output and residual values of the attack-free process over one simulation are illustrated in Fig. 4.3. The measured output values (Fig. 4.3a) and the residual values (Fig. 4.3b) of the process under both controller-observer parameters are maintained within their corresponding terminal set. Over the simulations, the switch into the attack-sensitive mode is implemented at a time step in the interval $[50, 56]$ ($[0.5, 0.56]$ h). At the time instance when the control system switches from the nominal mode to the attack-sensitive mode, the condition in Eq. 4.18 is satisfied over all simulations. As a result, this switch does not excite process dynamics. However, for the switch back to the nominal mode, the condition in Eq. 4.19 is not satisfied over the switching interval for all 1000 simulations, and the switch back to the nominal mode is forced at the end time $t_s + T_c^{max}$. Following this, alarms are suppressed for 2-time steps from the switch. No false alarms are observed because the augmented state converges to the attack-free minimum invariant set under nominal controller-observer parameters within 2-time steps or less from the switch. The results from one simulation are illustrated in Fig. 4.3. In this simulation, the monitoring variable values are contained within the attack-free terminal set under the nominal mode until the control system switches from the nominal mode to the attack-sensitive mode at the time $t_s = 52$ ($0.52\,\text{h}$). After the switch, the monitoring variable values are within the attack-free terminal sets under attack-sensitive controller-observer parameters (Fig. 4.3a and Fig. 4.3b). As a result, no false alarms are observed. Control system switches back to the nominal mode at the time $t_s^* = 162$ ($1.62\,\text{h}$). After the switch, the monitoring variable values are within its attack-free terminal set under nominal controller-observer parameters, and no alarms are observed.

The second scenario considers the attacked CSTR process with the active detection method to demonstrate the attack detection capabilities. A multiplicative attack of magnitude $\Lambda = \text{diag}(1, 0.85)$ is considered. The attack is potentially detectable under the nominal controller-observer parameters, and the attack is detectable under attack-sensitive controller-observer parameters. Over all simulations of the attacked process with the active detection method, the switch into attack sensitive mode is implemented over the time interval $[50, 74]$ ($[0.5, 0.74]$ h in real-time). The attack is detected in every simulation within 24 time steps after the switch into attack-sensitive mode. The results from one simulation are illustrated in Fig. 4.4. In this simulation, the attack is not detected with $\chi(t) \in D_\chi(I, K^*, L^*)$ for $t \in [0, t_s]$ (Fig. 4.4a, Fig. 4.4b). After the switch, the attack is detected at the time $t_d = 57$ (0.57 h) due to $\chi(t_d) \notin D_\chi(I, K_\Lambda, L_\Lambda)$ (Fig. 4.4a, Fig. 4.4b). Immediately after attack detection, the control system switches back to the nominal mode to stabilize the process. After the switch, the monitoring variable is contained within its attack-free terminal set under nominal controller-observer parameters and no further alarms are observed.

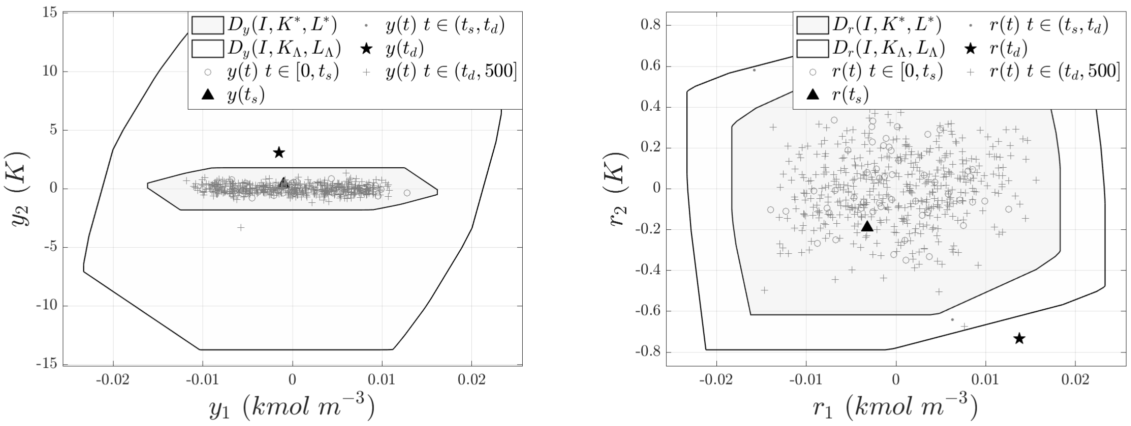


(a)

(b)

Fig. 4.3: (a) The output values over a simulation of the attack-free closed-loop process with the proposed active detection method. (b) The residual values over a simulation of the attack-free closed-loop process with the proposed active detection method.

For a basis of comparison, the closed-loop process is also simulated with the process operating exclusively under the nominal mode and monitored by the detection scheme in Eq. 4.14. In this case, the attack is detected in 20 out of 1000 simulations. The attack detection times over these simulations of the attacked process under nominal mode are compared with the attack detection times for the corresponding simulations of the attacked process with the active detection method. In 4 of the 20 simulations, the attack is detected before $t_i$. Over the corresponding 4 simulations with the active detection method, the attack is detected at the same time as the simulations of the process exclusively under the nominal mode. Over the remaining 16 of the 20 simulations of the process under the nominal mode, the attack is detected at a time in the interval $[80, 491]$ ($[0.8, 4.91]$ h). Over corresponding simulations of the process with the active detection method, the attack is detected at a time in the interval $[52, 64]$ ($[0.52, 0.64]$ h). Therefore, the active detection method enhances the detection capabilities of the detection scheme in Eq. 4.14.



(a)

(b)

Fig. 4.4: (a) The output values over a simulation of the attacked closed-loop process with the proposed active detection method. (b) The residual values over a simulation of the attacked closed-loop process with the proposed active detection method.

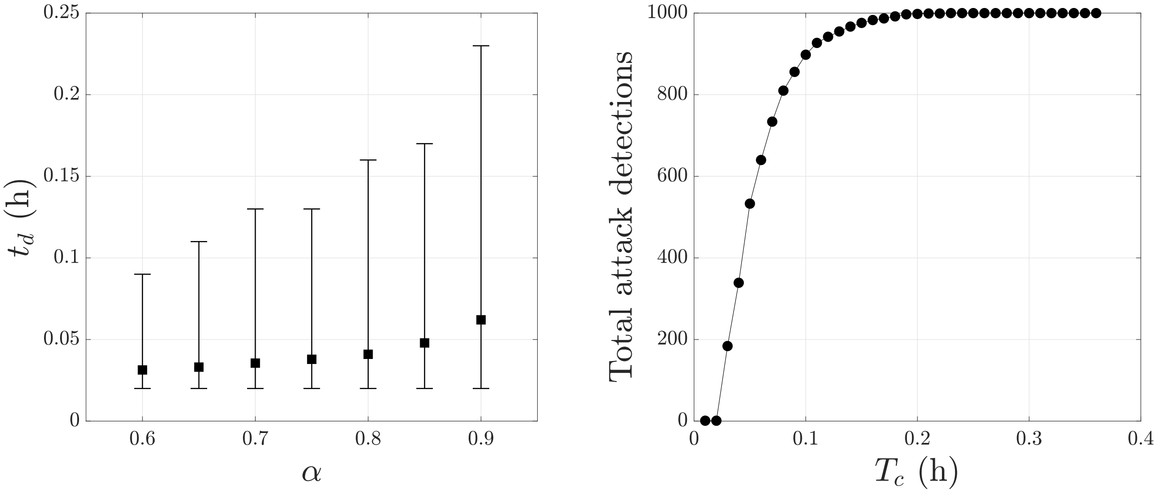
#### 4.4.2.1 Selection of a Minimum Dwell-Time for the CSTR Process

Using several simulations of the CSTR process under an attack, the choice of the minimum dwell-time of $T_c^{min} = 1\,\text{h}$ is analyzed. Several scenarios are considered. Each scenario consists of 1000 simulations of the CSTR process, similar to the scenarios in the prior section. To simulate the process in the attack-sensitive mode, the simulations are initialized with the attack-sensitive controller-observer parameters, i.e., for all scenarios considered, the switching time from the nominal to the attack-sensitive modes is $t_s = 0\,\text{h}$. A time-triggered switching strategy with a dwell-time of $T_c = 100$ under the attack-sensitive mode is used. Process states at each simulation are initialized at 0.

First, 7 different scenarios are considered to analyze if a minimum dwell-time of $T_c^{min} = 1\,\text{h}$ is sufficient to allow for the detection of attacks with magnitude in the range $\{\Lambda \mid \text{diag}(1, \alpha) \mid \alpha \in [0.6, 0.9]\}$. Across scenarios, the magnitude of attack targeting the temperature sensor-controller link is varied. The first scenario considers an attack of magnitude $\Lambda = \text{diag}(1, \alpha)$, with $\alpha = 0.6$. For each of the subsequent scenarios, $\alpha$ is incremented by 0.05 over the range until a value of $\alpha = 0.9$ is reached for the seventh scenario. The minimum, maximum and average time for detection of the attack are computed over each scenario, as illustrated in Fig. 4.5a. The average time for attack detection increases with the value of $\alpha$. The minimum detection time of all attacks is 0.03 h. The attack with $\alpha = 0.9$ has the maximum time for detection of $t_d = 0.23\,\text{h}$. Based on this result, a dwell-time of $T_c^{min} = 1\,\text{h}$ is sufficient to ensure the detection of attacks in the range $\{\Lambda \mid \text{diag}(1, \alpha) \mid \alpha \in [0.6, 0.9]\}$.

A second simulation study is conducted to analyze the impact of various dwell-times on attack detection. Several scenarios are considered for the process under an attack of magnitude with $\alpha = 0.9$. An attack with $\alpha = 0.9$ is chosen because it has the maximum detection time in the first simulation study. In total, 30 scenarios are considered. In the first scenario, a dwell-time of $T_c = 0.01\,\text{h}$ is chosen. Thereafter, for each scenario, the dwell-time is incremented by 0.01 h, with the last scenario considering a dwell-time of 0.3 h. Over each scenario, the total number of simulations out of 1000 simulations with an attack detection is computed (Fig. 4.5b). It is observed that as the dwell-time increases, the total attack detections also increase. Furthermore, a dwell-time of $T_c = 0.15\,\text{h}$ under the attack-sensitive controller-observer parameters may be sufficient to detect an attack in 97.6% of the simulations. Similarly, a dwell-time of $T_c = 0.23\,\text{h}$ in 100% of the simulations considered. Thereafter, a further increase in the dwell-time has no impact on the total attack detections. The results indicate that to limit the performance degradation in the process, a smaller dwell-time than $T_c^{min} = 1\,\text{h}$ may be considered.

**Remark 4.4.1.** *In this section, the proposed active detection method is applied to a nonlinear chemical process, extending beyond what is considered in the theoretical analysis presented in this work. From the closed-loop simulation results, the detection scheme detected the multiplicative attack, and did not raise any false alarms. Also, the augmented state is maintained within the minimum invariant set computed from the linearized process model in all cases. These results demonstrate the proposed active detection method's applicability to the nonlinear CSTR process. In general, it may be expected that the method* *will provide minimal false alarms while enhancing the detection capabilities for nonlinear processes when the augmented state is maintained in a small neighborhood of the origin such that the effect of the nonlinearities is small, i.e., when the process disturbances and measurement noise are small. However, extensions of the active detection method to nonlinear processes remain an open area and are subject to future work.*



(a)

(b)

Fig. 4.5: (a) The attack detection times for different attack magnitudes and a dwell-time of $T_c = 1\,\text{h}$. (b) The number of attacks detected under the attack-sensitive mode for an attack of magnitude of $\Lambda = \text{diag}(1, 0.9)$ with different dwell-times.

## 4.5 Conclusions

In this chapter, a detectability-based classification of multiplicative sensor-controller link false-data injection attacks with respect to a general class of detection schemes monitoring the process was presented. A control switching-based approach for enhancing attack detectability with respect to an output and residual-based detection scheme was proposed. To guarantee zero false alarms from switching, a confidence region for the attack-free augmented states was constructed, and a confidence region-based switching condition was developed. The switching condition was incorporated into the proposed active detection method to minimize false alarms. The application of the proposed active detection method for attack detectability enhancement and false alarm minimization was demonstrated using two illustrative processes.

# Chapter 5

# A Reachable Set-Based Scheme for the Detection of False Data Injection Cyberattacks on Dynamic Processes

In this section, a reachable set-based detection scheme is proposed to monitor transient process operations for false data injection attacks (FDIAs) that alter the variable value communicated over the PCS communication links. Both sensor-controller and controller-actuator link FDIAs are considered. The proposed detection scheme verifies whether the value of a generalized monitoring variable at a given time step is contained within its reachable set for the attack-free process. The proposed detection scheme monitors the process without requiring extensive closed-loop data. It also does not raise false alarms during transient operation. Conditions that lead to an attack being detectable or undetectable with respect to the proposed detection scheme are characterized. The proposed detection scheme and the classification approach are applied to two illustrative process examples. The detectability of different FDIAs is analyzed, and the applicability of the reachable set-based detection scheme and attack classification to a nonlinear chemical process is demonstrated.

## 5.1 Preliminaries

### 5.1.1 Notation and Definitions

The set of non-negative integers is denoted by $\mathbb{Z}^+$. Given $\mathcal{W} \subseteq \mathbb{R}^n$ and $\mathcal{V} \subseteq \mathbb{R}^n$, the Minkowski sum of $\mathcal{W}$ and $\mathcal{V}$ is given by $\mathcal{W} \oplus \mathcal{V} := \{w + v \mid w \in \mathcal{W}, v \in \mathcal{V}\}$. For the set $\mathcal{D} \subseteq \mathbb{R}^n$ and matrix $\mathbb{R}^{m \times n}$, $A\mathcal{D} := \{Ad \mid d \in \mathcal{D}\}$. For a square matrix $A$, $\lambda_i(A)$ represents the $i^{\text{th}}$ eigenvalue of $A$. The identity matrix is denoted by $I$.

### 5.1.2 Class of Attack-Free Processes

Discrete-time linear processes with the following state-space dynamics are considered:

$$x_{k+1} = Ax_k + B^u u_k + B^w w_k \tag{5.1}$$

where $A \in \mathbb{R}^{n_x \times n_x}$, $B^u \in \mathbb{R}^{n_x \times n_u}$, $B^w \in \mathbb{R}^{n_x \times n_w}$, $k \in \mathbb{Z}^+$ is the time step, $x_k \in \mathbb{R}^{n_x}$ is the state vector, $u_k \in \mathbb{R}^{n_u}$ is the manipulated input vector, and $w_k \in \mathcal{W} \subset \mathbb{R}^{n_w}$ is the process disturbance vector. Without loss of generality, the initial time step is assumed to be $k = 0$. Measurements from the process are available and are given by:

$$y_k = Cx_k + v_k \tag{5.2}$$

where $y_k \in \mathbb{R}^{n_y}$ is the measured output vector and $v_k \in \mathcal{V} \subset \mathbb{R}^{n_y}$ is the measurement noise vector. The sets $\mathcal{W}$ and $\mathcal{V}$ are the sets of admissible process disturbances and measurement noise, respectively, and are assumed to be convex polytopes. A Luenberger observer is synthesized to compute state estimates as follows:

$$\begin{aligned} \hat{x}_{k+1} &= A\hat{x}_k + B^u u_k + L(y_k - \hat{y}_k) \\ \hat{y}_k &= C\hat{x}_k \end{aligned} \tag{5.3}$$

where $L \in \mathbb{R}^{n_x \times n_y}$ is the observer gain, $\hat{x}_k \in \mathbb{R}^{n_x}$ is the estimated state, and $\hat{y}_k \in \mathbb{R}^{n_y}$ is the estimated output. The estimation error, defined as the difference between the process state and the estimate ($e_k := x_k - \hat{x}_k$), has the following dynamics:

$$e_{k+1} = (A - LC)e_k + B^w w_k - Lv_k \tag{5.4}$$

The observer gain $L$ is selected such that all eigenvalues of the matrix $A - LC$ lie within the unit circle. The control objective is to stabilize the closed-loop process around its steady-state, assumed to be the origin of the unperturbed system. To achieve the control objective, a linear control law of the following form is synthesized:

$$u_k = -K\hat{x}_k \tag{5.5}$$

where $K \in \mathbb{R}^{n_u \times n_x}$ is the controller gain. The controller gain $K$ is selected to ensure that all eigenvalues of the matrix $A - BK$ lie within the unit circle.

The dynamics of the process state and the estimation error collectively capture the attack-free closed-loop process dynamics. An augmented state vector is defined and denoted by $\xi_k := [x_k^T \ e_k^T]^T$, with dynamics:

$$\xi_{k+1} = \underbrace{\begin{bmatrix} A - B^u K & B^u K \\ 0 & A - LC \end{bmatrix}}_{=:A^\xi} \xi_k + \underbrace{\begin{bmatrix} B^w & 0 \\ B^w & -L \end{bmatrix}}_{=:B^d} d_k \tag{5.6}$$

where $d_k \in \mathcal{D} := \mathcal{W} \times \mathcal{V}$ is a concatenated vector that includes the process disturbance and measurement noise vectors, i.e., $d_k := [w_k^T \ v_k^T]^T$. The input $d_k$ is called the disturbance for simplicity. The augmented system described in Eq. 5.6 is referred to as the attack-free closed-loop process. For the closed-loop process, its initial set $\mathcal{R}_0^\xi \subset \mathbb{R}^{2n_x}$ is defined as the region in state space that contains the value of the augmented state at time step $k = 0$, i.e., $\xi_0 \in \mathcal{R}_0^\xi \subset \mathbb{R}^{2n_x}$. The initial set is assumed to be a polytope. Provided a set of initial states $\mathcal{R}_0^\xi \subset \mathbb{R}^{2n_x}$, the $k$-step forward reachable set, denoted by $\mathcal{R}_k^\xi(\mathcal{R}_0^\xi)$, for the closed-loop process is the set consisting of all states that can be reached in $k$ time steps under any admissible disturbance, and is given by (e.g., the unlabeled equation preceding Eq. 2 in [90]):

$$\mathcal{R}_k^\xi(\mathcal{R}_0^\xi) = {A^\xi}^k \mathcal{R}_0^\xi \bigoplus_{i=0}^{k-1} {A^\xi}^i B^d \mathcal{D} \tag{5.7}$$

The $k$-step reachable set for the attack-free closed-loop process depends on the controller-observer gains $(K, L)$, the disturbance set $\mathcal{D}$, and the initial set $\mathcal{R}_0^\xi$. As $k \to \infty$, the $k$-step forward reachable sets converge to the minimum invariant set ($\mathcal{R}_\infty^\xi := \bigoplus_{i=0}^\infty {A^\xi}^i B^d \mathcal{D}$), which is the limit set for all trajectories of the process ([79]).

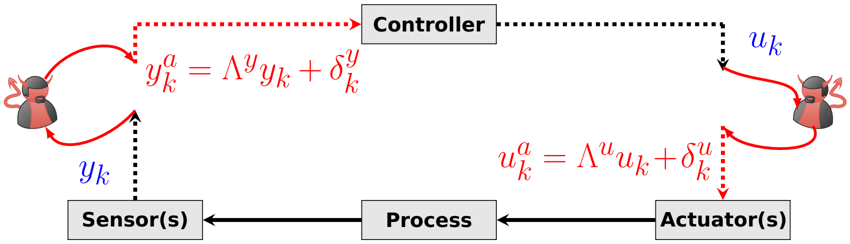

Controller
$u_k$
$y_k^a = \Lambda^y y_k + \delta_k^y$
$u_k^a = \Lambda^u u_k + \delta_k^u$
$y_k$
Sensor(s)
Process
Actuator(s)

Fig. 5.1: A block diagram illustrating a process control system under a false data injection attack that simultaneously alters the data over sensor-controller and controller-actuator communication links.

**Remark 5.1.1.** *The initial set $\mathcal{R}_0^\xi$ and the disturbance set $\mathcal{D}$ are assumed to be polytopes. With these assumptions, Eq. 7 can be computed by recursively applying the following two properties: (1) for two polytopes $\mathcal{D}_1$ and $\mathcal{D}_2$, $\mathcal{D}_1 \oplus \mathcal{D}_2$ can be computed by adding the vertices of $\mathcal{D}_1$ to the vertices of $\mathcal{D}_2$ where the resulting vectors form the vertices of $\mathcal{D}_1 \oplus \mathcal{D}_2$, and (2) for a polytope $\mathcal{D}$, $A^{\xi^i} B^d \mathcal{D}$ is a polytope that can be computed by pre-multiplying all vertices of $\mathcal{D}$ by $A^{\xi^i} B^d$ and taking the convex hull of the resultant vectors. The assumption that $\mathcal{R}_0^\xi$ and $\mathcal{D}$ are polytopes enables the calculation of Eq. 7 with a finite number of computations.*

### 5.1.3 False Data Injection Attacks

False data injection attacks (FDIAs) refer to cyberattacks that alter the state or input values communicated over a communication link so that the receiver, i.e., the controller or the actuators, receives the altered value. In the present work, both additive and multiplicative FDIAs that alter data communicated over the sensor-controller and controller-actuator communication links are considered. In the presence of an attack, the value of the variable altered by the attack is given by:

$$\phi_k^a = \Lambda^\phi \phi_k + \delta_k^\phi \tag{5.8}$$

where $\phi_k \in \mathbb{R}^{n_\phi}$ is the unaltered value of the variable, $\phi_k^a$ is the altered value of the variable $\phi_k$, $\Lambda^\phi \in \mathbb{R}^{n_\phi \times n_\phi}$ is a multiplicative factor to represent multiplicative FDIAs, and $\delta_k^\phi \in \mathbb{R}_k^{n_\phi}$ is the additive bias to represent additive FDIAs. For sensor-controller link FDIAs, $\phi_k$ represents the sensor measurements ($y_k$); for controller-actuator link FDIAs, $\phi_k$ represents the controller output ($u_k$). Figure 5.1 illustrates the block diagram of a process control system under a false data injection attack that simultaneously alters the data over the sensor-controller and controller-actuator links. In the presence of an attack, the values of the measured output and the control input ($y_k$ and $u_k$ shown in blue text) are altered by the attacker and reported over the compromised communication links as $y_k^a = \Lambda^y y_k + \delta_k^y$ and $u_k^a = \Lambda^u u_k + \delta_k^u$, respectively (shown in red text).

FDIAs alter the closed-loop behavior of the process. The augmented state dynamics of the closed-loop process subject to an additive and multiplicative FDIA are given by:

$$\xi_{k+1} = \underbrace{\begin{bmatrix} A - B^u\Lambda^u K & B^u\Lambda^u K \\ L(I-\Lambda^y)C & A - LC \end{bmatrix}}_{=:A^{\xi_a}} \xi_k + \underbrace{\begin{bmatrix} B^w & 0 \\ B^w & -L\Lambda^y \end{bmatrix}}_{=:B^{d_a}} d_k + \underbrace{\begin{bmatrix} 0 & B^u \\ -L & 0 \end{bmatrix}}_{=:B^{\delta_a}} \delta_k \tag{5.9}$$

where $\delta_k = \begin{bmatrix} (\delta_k^y)^T & (\delta_k^u)^T \end{bmatrix}^T$. The closed-loop process described by Eq. 5.9 is referred to as the attacked closed-loop process. Similar to the attack-free process, the $k$-step reachable set under an FDIA is given by:

$$\mathcal{R}_k^{\xi_a}(\mathcal{R}_0^{\xi}) = {A^{\xi_a}}^k \mathcal{R}_0^{\xi} \bigoplus_{i=0}^{k-1} {A^{\xi_a}}^i \mathcal{D}_k^a \tag{5.10}$$

where $\mathcal{D}_k^a = B^{d_a}\mathcal{D} \oplus B^{\delta_a}\{\delta_k\}$. The attack is generally unknown, so the $k$-step reachable sets of the attacked process may not be computable for purposes of online attack detection. However, the $k$-step reachable sets of the attacked process can be used for (offline) classification of specific attacks as detectable or not (this point is discussed further in Section 5.2.2).

## 5.2 Attack Detection for Processes During Transient Operation

In this section, a class of reachable set-based attack detection schemes utilizing a generalized monitoring variable are presented to monitor the closed-loop process during transient operation. A method for classifying attacks as detectable, potentially detectable, or undetectable under the proposed detection scheme is also presented.

### 5.2.1 Reachable Set-Based Detection Scheme

Cyberattack detection schemes often use the measured output, estimated output, or the residual vector ($r_k \coloneqq y_k - \hat{y}_k$) as the monitoring variable(s) to detect an attack (e.g., [23, 50, 53, 91, 92]). Some attacks may evade detection by a scheme that uses only one of the three variables, but may be detected using a detection scheme based on another variable ([48]). In this work, a generalized monitoring variable that may be expressed as a linear combination of the measured output and its estimate generated by the observer is considered:

$$\eta_k := H^y y_k + H^{\hat{y}} \hat{y}_k \tag{5.11}$$

where $\eta_k \in \mathbb{R}^{n_\eta}$ is the generalized monitoring variable, and the matrices $H^y$ and $H^{\hat{y}}$ are design parameters of the detection scheme. When $H^y$ and $H^{\hat{y}}$ are chosen such that $H^y = I$ and $H^{\hat{y}} = -I$, the monitoring variable becomes the residual vector ($\eta_k = r_k$). A choice of $H^y = I$ and $H^{\hat{y}} = 0$, on the other hand, results in the monitoring variable being the measured output. Expressing the monitoring variable in terms of the augmented state and the disturbance vector gives:

$$\eta_k = \underbrace{\begin{bmatrix}(H^y - H^{\hat{y}})C & H^{\hat{y}}C\end{bmatrix}}_{=:C^\xi} \xi_k + \underbrace{\begin{bmatrix}0 & H^y\end{bmatrix}}_{=:D^d} d_k \tag{5.12}$$

To address the problem of attack detection during transient operation, the reachable sets of the monitoring variable for the attack-free closed-loop process are considered. For the attack-free closed-loop process and initial set $\mathcal{R}_0^\xi$, the augmented state is contained within the $k$-step reachable set for all $k \in \mathbb{Z}^+$. From Eqs. 5.7 and 5.12, the generalized monitoring variable of the attack-free process is contained in the set:

$$\mathcal{R}_k^\eta(\mathcal{R}_k^\xi) := C^\xi \mathcal{R}_k^\xi(\mathcal{R}_0^\xi) \oplus D^d \mathcal{D} \tag{5.13}$$

The containment of the monitoring variable within the $k$-step reachable sets of the attack-free process may be verified to monitor the process for attacks as follows:

$$h(\eta_k, \mathcal{R}_k^\xi) = \begin{cases} 1, & \eta_k \notin \mathcal{R}_k^\eta(\mathcal{R}_k^\xi) \\ 0, & \eta_k \in \mathcal{R}_k^\eta(\mathcal{R}_k^\xi) \end{cases} \tag{5.14}$$

where the mapping $h$ returns the output of the detection scheme. An output of 1 indicates that an attack is detected, and the detection scheme is said to raise an alarm. An output of 0 indicates that no attack is detected. To implement the reachable set-based detection scheme, knowledge of the initial set is required. For a process transitioning from one steady-state to another, the minimum invariant set of the process at the initial steady-state may be used as the initial set. The initial set for process start-up may also be known.

The reachable set-based detection scheme in Eq. 5.14 is designed to detect an attack if there is a discrepancy between the observed value of the monitoring variable and its expected attack-free value, i.e., the reachable sets are computed for the attack-free process. In the absence of an attack, the values of the generalized monitoring variable are contained within $k$-step reachable sets for the attack-free process, and the detection scheme generates an output of 0 for all $k \in \mathbb{Z}^+$. Therefore, a necessary condition for attack-free operation is that the monitoring variable must be contained within its $k$-step reachable set, implying that no attacks are detected. This is formalized in the following proposition.

**Proposition 8.** *Consider the closed-loop process in Eq. 5.9 monitored by the reachable set-based detection scheme in Eq. 5.14, with an initial set $\mathcal{R}_0^{\xi}$. The closed-loop process is attack-free only if the output of the detection scheme in Eq. 5.14 is $h(\eta_k, \mathcal{R}_k^{\xi}) = 0$ for all $k \in \mathbb{Z}^+$.*

A direct implication of Proposition 8 is that the detection scheme does not raise false alarms during transient operation. While the reachable set-based detection scheme is designed on the basis of attack-free process behavior, an attack may be detected if the detection scheme returns a value of 1 at some $k \in \mathbb{Z}^+$.

**Corollary 2.** *Consider the closed-loop process in Eq. 5.9 monitored by the reachable set-based detection scheme in Eq. 5.14, with an initial set $\mathcal{R}_0^{\xi}$. If the output of the detection scheme is $h(\eta_{k_d}, \mathcal{R}_{k_d}^{\xi}) = 1$ for some $k_d \in \mathbb{Z}^+$, then the process cannot be attack-free.*

**Remark 5.2.1.** *With respect to the reachable set-based detection scheme, the detectability of an attack depends on how the reachable sets of the monitoring variable for the attacked process evolve with respect to the evolution of the reachable sets of the monitoring variable for the attack-free process. The detectability-based classification of an attack may vary with the monitoring variable (i.e., the choice of $H^y$ and $H^{\hat{y}}$). From Eq. 5.14, the parameters $H^y$*

*and $H^{\hat{y}}$ influence the reachable sets of the monitoring variable for the attack-free process. For the attacked process evolving from an initial set $\mathcal{R}_0^{\xi} \subset \mathbb{R}^{2n_x}$, the reachable sets of the monitoring variable are also influenced by the parameters $H^y$ and $H^{\hat{y}}$:*

$$\mathcal{R}_k^{\eta_a}(\mathcal{R}_k^{\xi_a}) = C^{\xi_a}\mathcal{R}_k^{\xi} \oplus \mathcal{D}^{\eta_a} \tag{5.15}$$

*where $C^{\xi_a} = \begin{bmatrix}(H^y\Lambda^y + H^{\hat{y}})C & -H^{\hat{y}}C\end{bmatrix}$, $\mathcal{D}^{\eta_a} = \begin{bmatrix}0 & H^y\Lambda^y\end{bmatrix} d_k \oplus \begin{bmatrix}H^y & 0\end{bmatrix}\{\delta_k\}$. Based on Eq. 5.14 and Eq. 5.15, the parameters $H^y$ and $H^{\hat{y}}$ influence the evolution of the reachable sets of the monitoring variable for the attacked and the attack-free process. Therefore, $H^y$ and $H^{\hat{y}}$ influence attack detectability.*

### 5.2.2 Classification of Attack Detectability

Attacks can be classified based on the ability or inability of the reachable set-based detection scheme to detect an attack. Defining attack detectability requires certain considerations, including the dependence of reachable sets on the initial set $\mathcal{R}_0^{\xi}$. An attack is detected at time $k_d$ if $h(\eta_{k_d}, \mathcal{R}_{k_d}^{\xi}) = 1$. An attack is detectable with respect to the reachable set-based detection scheme and the initial set $\mathcal{R}_0^{\xi}$ if the attack is detected in finite time for all $\xi_0 \in \mathcal{R}_0^{\xi}$ (and $d_k \in \mathcal{D}$). An attack is undetectable with respect to the reachable set-based detection scheme and the initial set $\mathcal{R}_0^{\xi}$ if the attack is not detected in finite time for all $\xi_0 \in \mathcal{R}_0^{\xi}$ (and $d_k \in \mathcal{D}$). For simplicity of presentation, detectable and undetectable attacks with respect to the detection scheme and initial set $\mathcal{R}_0^{\xi}$ are called detectable and undetectable attacks, respectively. An attack is potentially detectable if it is neither detectable nor undetectable.

With the definitions above, conditions based on the relationship between the reachable sets of the attacked and the attack-free process can be established and used for classifying attacks. In the propositions below, an FDIA that begins at $k = 0$ is considered. The results may be extended to an attack occurring at any time. The first proposition establishes that if all possible values of the monitoring variable of the attacked process are contained within the reachable sets for the attack-free process, then the attack is undetectable.

**Proposition 9.** *Consider the closed-loop process in Eq. 5.9, with an initial set $\mathcal{R}_0^{\xi}$, under an FDIA beginning at $k = 0$. The attack is undetectable with respect to the detection scheme in Eq. 5.14 and the initial set $\mathcal{R}_0^{\xi}$ if and only if the reachable sets of the monitoring variable for the attacked and the attack-free process satisfy $\mathcal{R}_k^{\eta_a}(\mathcal{R}_k^{\xi_a}) \subseteq \mathcal{R}_k^{\eta}(\mathcal{R}_k^{\xi})$ for all $k \in \mathbb{Z}^+$.*

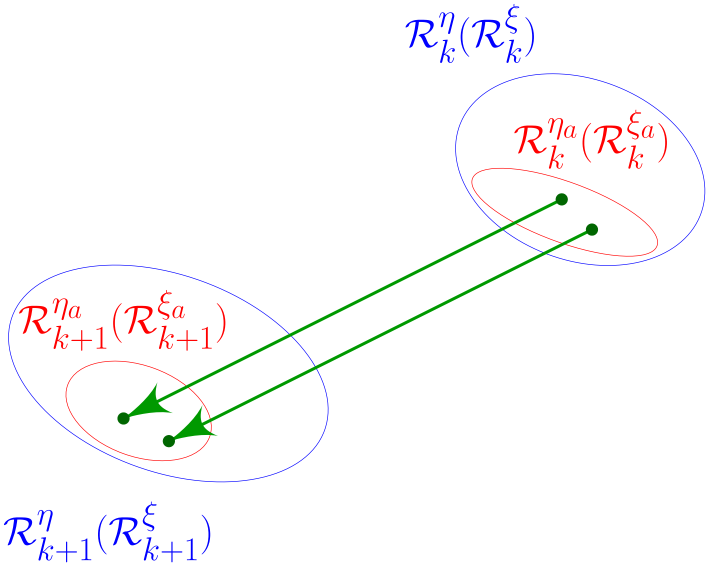

$\mathcal{R}_k^{\eta}(\mathcal{R}_k^{\xi})$
$\mathcal{R}_k^{\eta_a}(\mathcal{R}_k^{\xi_a})$
$\mathcal{R}_{k+1}^{\eta_a}(\mathcal{R}_{k+1}^{\xi_a})$
$\mathcal{R}_{k+1}^{\eta}(\mathcal{R}_{k+1}^{\xi})$

Fig. 5.2: Illustrative example showing the reachable sets of the monitoring variable for the attack-free (blue sets) and the attacked (red sets) process in the presence of an undetectable attack, with two example trajectories (green lines) for the attacked process.

The condition presented in Proposition 9 is a necessary and sufficient condition for an undetectable attack. Figure 5.2 provides a pictorial interpretation of the result of Proposition 9. It illustrates the reachable sets of the attacked process (sets in red) and the attack-free process (sets in blue) over two time steps for a process under an undetectable attack. The figure also illustrates two example trajectories of the monitoring variable for the attacked process (green lines). As illustrated, the values of the monitoring variable at the time steps $k$ and $k + 1$ (shown by the green circle markers) are contained within the intersection of the reachable sets for the attack-free and attacked process, leading to an output of 0 from the detection scheme. Therefore, the attack is not detected by the detection scheme. While only two time steps are illustrated in Fig. 5.2, the reachable sets of the process under an undetectable attack must be contained within the attack-free reachable sets for all time steps $k \in \mathbb{Z}^+$.

If the reachable set of the monitoring variable for the attacked process does not intersect the reachable set of the attack-free process at some time $k \in \mathbb{Z}^+$, the attack will be detected at time $k$, and is detectable. This is formally stated in the following proposition.

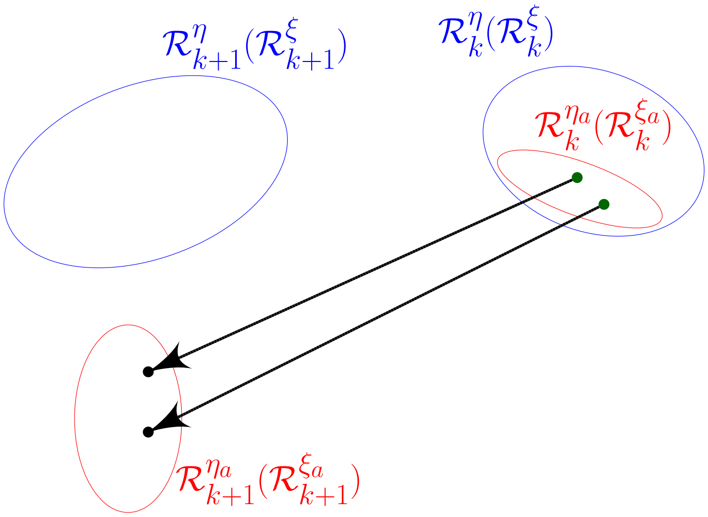

$\mathcal{R}^{\eta}_{k+1}(\mathcal{R}^{\xi}_{k+1})$
$\mathcal{R}^{\eta}_{k}(\mathcal{R}^{\xi}_{k})$
$\mathcal{R}^{\eta_a}_{k}(\mathcal{R}^{\xi_a}_{k})$
$\mathcal{R}^{\eta_a}_{k+1}(\mathcal{R}^{\xi_a}_{k+1})$

Fig. 5.3: Illustrative example showing the reachable sets of the monitoring variable for the attack-free (blue sets) and the attacked (red sets) process in the presence of a detectable attack, with two example trajectories for the attacked process.

**Proposition 10.** *Consider the closed-loop process in Eq. 5.9, with an initial set* $\mathcal{R}^{\xi}_0$, *under an FDIA beginning at* $k = 0$. *The attack is detectable if the reachable sets of the monitoring variable for the attacked and the attack-free process satisfy* $\mathcal{R}^{\eta_a}_k(\mathcal{R}^{\xi_a}_k) \cap \mathcal{R}^{\eta}_k(\mathcal{R}^{\xi}_k) = \emptyset$ *for some* $k \in \mathbb{Z}^+$.

Fig. 5.3 provides an illustration of the idea behind Proposition 10. The figure shows the reachable sets of the monitoring variable for the attack-free process (blue sets) and those of the process under a detectable attack (red sets) over two time steps. At time step $k$, the reachable set for the attacked process is contained entirely within the reachable set for the attack-free process. At time step $k + 1$, the reachable set of the attacked process does not intersect the reachable set of the attack-free process. For all initial values, no value of the monitoring variable of the attacked process is contained in the reachable set of the attack-free process (illustrated by the black circle markers in Fig. 5.3). As a result, the attack is detected at time step $k+1$ with the detection scheme generating an output of 1, i.e., $h(\eta_{k+1}, \mathcal{R}^{\xi}_{k+1}) = 1$ for all $\xi_0 \in \mathcal{R}^{\xi}_0$.

Attacks that do not satisfy the conditions in Proposition 9 or Proposition 10 are also possible. For such an attack, the reachable sets of the attacked process intersect with the reachable sets of the attack-free process for all time steps, and the reachable sets of the attacked process are not contained in the corresponding reachable sets of the attack-free process for at least one time step, i.e., $\mathcal{R}^{\eta}_k(\mathcal{R}^{\xi}_k) \cap \mathcal{R}^{\eta_a}_k(\mathcal{R}^{\xi_a}_k) \neq \emptyset$ for all $k \in \mathbb{Z}^+$ and $\mathcal{R}^{\eta_a}_k(\mathcal{R}^{\xi_a}_k) \not\subseteq \mathcal{R}^{\eta}_k(\mathcal{R}^{\xi}_k)$ for some $k \in \mathbb{Z}^+$. While the attack cannot be undetectable by Proposition 9, the attack may be detectable or potentially detectable. For example, consider an attack on the process such that the monitoring variable of all possible trajectories leaves its attack-free reachable set. In this case, the attack is detectable. This is illustrated by the following example:

$$\begin{aligned} x_{k+1} &= 0.9x_k + \delta^u_k \\ y_k &= x_k \end{aligned} \tag{5.16}$$

where $x_k \in \mathbb{R}$ is the state, $y_k \in \mathbb{R}$ is the measurement, and $\delta^u_k \in \mathbb{R}$ is an additive controller-actuator link FDIA. Consider the initial set of $\{0\}$ and let the measured output be the monitoring variable, i.e., $\eta_k = y_k$. For the attack-free process, the monitoring variable takes a value of 0 for all $k$, and the reachable sets of the monitoring variable are $\{0\}$ for all $k \in \mathbb{Z}^+$. Let the attack $\delta^u_k$ be a bounded random variable such that $|\delta^u_k| \leq \bar{\delta}$ for all $k \in \mathbb{Z}^+$, where $\bar{\delta} > 0$. Moreover, let $\delta^u_k$ take a non-zero value for at least one time step. The reachable sets of the attacked process contain the origin, so they intersect with the attack-free reachable sets for all $k \in \mathbb{Z}^+$. When $\delta^u_k$ takes a non-zero value, the state and monitoring variable will move away from 0, so the attack will be detected. Thus, the attack is detectable.

An attack that does not satisfy the conditions in Proposition 9 or Proposition 10 may be potentially detectable if there are some trajectories where the attack is detected and others where the attack is not detected. For example, consider the following process:

$$\begin{aligned} x_{k+1} &= 0.9x_k + d_k \\ y_k &= \Lambda^y x_k \end{aligned} \tag{5.17}$$

where $d_k$ is the process disturbance taking values in the set $[-1, 1]$, and $\Lambda^y = 1.1$ is a multiplicative FDIA altering the data over the sensor-controller link. For the attack-free process with $A^\xi = 0.9$, $B^d = 1$, and disturbances bounded as $\mathcal{D} = [-1, 1]$, the minimum invariant set (computed based on the method presented in [79]) is $[-10, 10]$, meaning that for any $d_k \in \mathcal{D}$, $x_{k+1} \in [-10, 10]$ if $x_k \in [-10, 10]$. Let the initial set be equal to the minimum invariant set of the attack-free process, i.e., $[-10, 10]$, and let the monitoring variable be the measured output. For a process evolving from an initial set $\mathcal{R}_0^\xi \in [-10, 10]$, the $k$-step forward reachable set of the process is the minimum invariant set itself. Therefore, the reachable sets of the monitoring variable are $[-10, 10]$ for all $k \in \mathbb{Z}^+$. If the initial state of the attacked process is $x_0 = 0$ and the disturbance takes a value of zero for all time, i.e., $d_k = 0$ for all $k \in \mathbb{Z}^+$, the monitoring variable of the attacked process takes a value of 0 for all time, and the attack will not be detected. For some other initial states and disturbances, the attack will be detected. If $x_0 = 10$ and $d_0 = 1$, for example, the value of the monitoring variable is not contained within the reachable set at $k = 1$, since $\eta_1 = 11 \notin [-10, 10]$. In this case, the attack is detected. The attack is potentially detectable because there are some trajectories for which the attack will be detected and other trajectories for which the attack is not detected.

**Remark 5.2.2.** *From Eq. 5.10, the reachable sets of the attacked process depend on the initial state and the matrices $A^{\xi_a}$, $B^{d_a}$, and $B^{\delta_a}$, which depend on the controller-observer gains (from Eq. 5.10). Therefore, the detectability of an attack is influenced by the controller-observer gains and the initial set. The dependence of attack detectability on the controller-observer gains and the initial set may be exploited to design methods that help attack detection.*

**Remark 5.2.3.** *An additional factor that may influence attack detectability is the closed-loop stability of the attacked process. Specifically, when the magnitudes of the multiplicative components of an attack ($\Lambda^y$ and $\Lambda^u$) are such that $\max_i |\lambda_i(A^{\xi_a})| \geq 1$, the attack destabilizes the process and may cause an unbounded growth in the norm of the augmented state. If an additional observability condition is satisfied ([50]), the attack may be detected because the generalized monitoring variable may not be contained within its k-step reachable* *set for the attack-free process at some time step $k \in \mathbb{Z}^+$.*

**Remark 5.2.4.** *For the attacked closed-loop process, the computation of the k-step reachable sets requires knowledge of the attack, which is unknown in general. Therefore, the detectability-based classification of attacks may be performed (offline) for various attacks.*

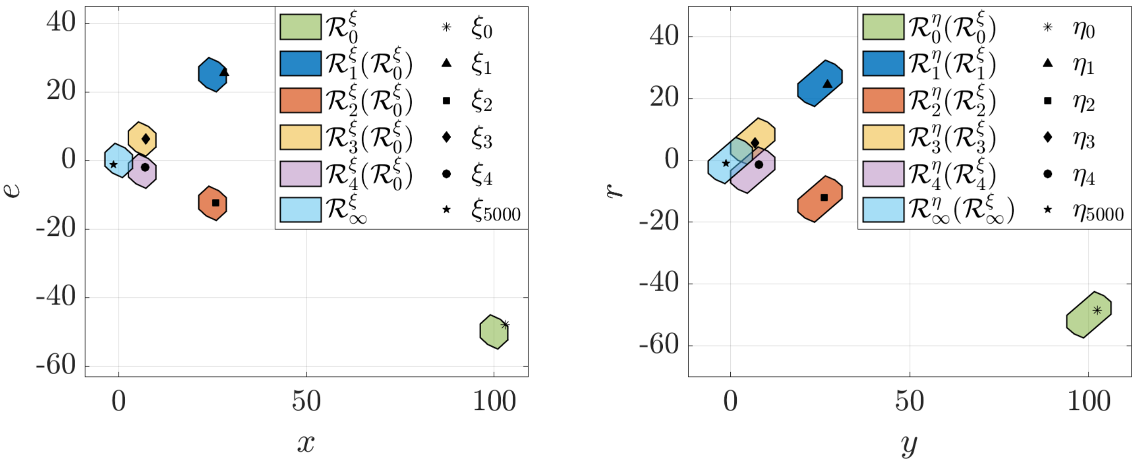
## 5.3 Numerical Results: Scalar Process Example

In this section, the proposed reachable set-based detection scheme, and the detectability-based classification of attacks, are applied to a scalar process during transient operation. All polytope computations are performed using the MPT 3.0 toolbox ([82]). A scalar process with the following process dynamics, measurement output, and control action is considered:

$$\begin{aligned} x_{k+1} &= x_k + u_k + w_k \\ u_k &= -\Lambda^u K \hat{x}_k + \delta_k^u \\ y_k &= \Lambda^y (x_k + v_k) + \delta_k^y \end{aligned}$$

where $x_k \in \mathbb{R}$ is the state, $u_k \in \mathbb{R}$ is the control action received by the actuator, $w_k \in \mathcal{W} := \{w' \mid |w'| \leq 1\}$ is the process disturbance, $y_k \in \mathbb{R}$ is the measurement output received by the controller, and $v_k \in \mathcal{V} := \{v' \mid |v'| \leq 1\}$ is the measurement noise. The process disturbance and measurement noise are modeled as random variables following a uniform distribution bounded between $-1$ and 1. The process may be subject to an FDIA that simultaneously alters the data communicated over the controller-actuator and the sensor-controller links. To monitor the process for attacks, a monitoring variable that is a concatenation of the measured output and the residual vector is chosen, i.e., $\eta_k = [y_k \; r_k]^T$. The monitoring variable fits the model for the generalized monitoring variable in Eq. 5.11 with $H^y = [1 \; 1]^T$ and $H^{\hat{y}} = [0 \; -1]^T$.

The process evolving from an initial set to the minimum invariant set is considered. A detection scheme tuned for steady-state operation (e.g., the detection scheme presented in [50]) is not applicable to monitor the process because it may raise alarms as the process evolves from its initial condition to the minimum invariant set during attack-free operation. Instead, the reachable set-based detection scheme in Eq. 5.14 is utilized. Three case studies are presented in this section. Each case study considers the process under a different attack. In the first case study, the application of the reachable set-based detection scheme is demonstrated. Additionally, the detectability-based classification of a simultaneous additive and multiplicative FDIA, which alters the data over the sensor-controller and controller-actuator links, is presented. In the second and third case studies, an additive FDIA and a multiplicative FDIA are considered, respectively. In all cases below, the polytope representing the initial set considered is the the attack-free minimum invariant set by shifted a vector (i.e., $\mathcal{R}_0^\xi = \mathcal{R}_\infty^\xi \oplus \{\xi'\}$ where $\xi'$ is the shifting vector).



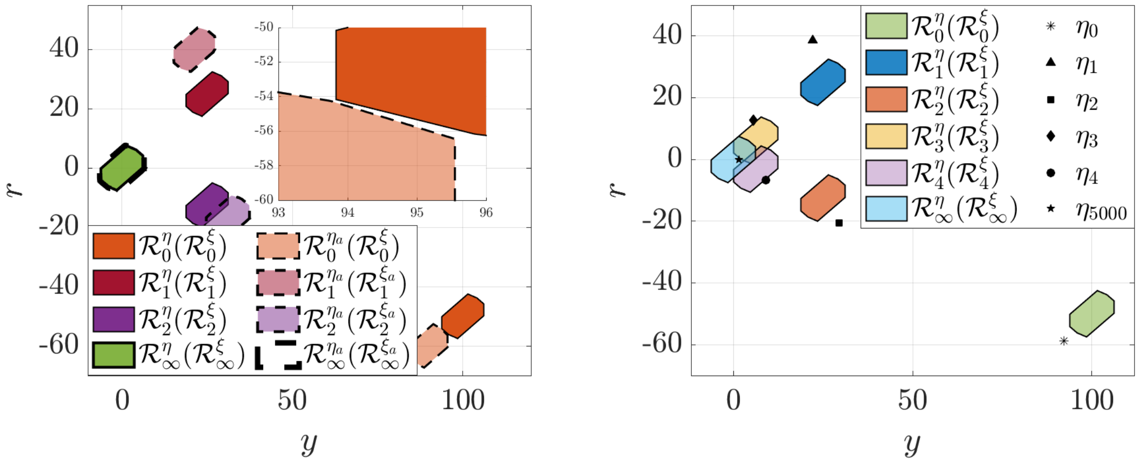
(a)

(b)

Fig. 5.4: (a) The state and estimation error values, (b) the monitoring variable values, used in the reachable set-based scheme, and (a)-(b) their corresponding reachable sets for the attack-free process over five time steps.

### 5.3.1 Application of the Reachable Set-Based Detection Scheme and Detectability-Based Classification of Attacks

The process evolving from an initial set that is the attack-free minimum invariant set shifted by $\xi' = [100 \;\; -50]^T$ ($\mathcal{R}_0^\xi$ in Fig. 5.4a) is considered. The attack-free process (first simulation set) and attacked process (second simulation set) are considered to demonstrate the reachable set-based detection scheme. Each simulation set consists of 1000 simulations of the process evolving from $\xi_0 = [103 \; -48]^T \in \mathcal{R}_0^\xi$ to the minimum invariant set. The total length of each simulation is 5000 time steps. For the attacked process, the cyberattack begins at $k = 0$, and is an FDIA with multiplicative factors $\Lambda^y = 0.9$ and $\Lambda^u = 1.05$ and additive biases, which are random variables drawn from a uniform distribution, where $\delta_k^y \in [-0.1, 0.1]$ and $\delta_k^u \in [-0.1, 0.1]$ for all $k \in \mathbb{Z}^+$. For demonstration purposes, the controller-observer gains are chosen as $K = 0.5$ and $L = 1.5$ because the attack on the process operated with $K = 0.5$ and $L = 1.5$ is found to be detectable, as described below.



(a) (b)

Fig. 5.5: (a) The reachable sets of the monitoring variable for the attack-free and the attacked process over a few time steps. At $k = 0$ and $k = 1$, $\mathcal{R}_k^{\eta} \cap \mathcal{R}_k^{\eta_a} = \emptyset$, indicating the attack is detectable. The localized zoom in the figure illustrates that the reachable sets at $k = 0$ do not intersect. (b) The values of the monitoring variable of the attacked process and the reachable sets used in the detection scheme over a few time steps.

In the first simulation set, the attack-free process is considered. In every simulation, the values of the state and the estimation error are contained within the reachable sets of the attack-free process. Similarly, the values of the monitoring variable are always contained within their reachable sets. Therefore, the output of the reachable set-based detection scheme is equal to 0 in all simulations, indicating a lack of attack detection. Figure 5.4a illustrates the values of the state and estimation error over one simulation of the attack-free process, and Fig. 5.4b illustrates the values of the output and estimation error over the same simulation. The values of all variables are contained within their corresponding reachable sets over the simulation, and no alarms are raised. The result demonstrates that the reachable set-based detection scheme does not raise false alarms during dynamic operation.

The second simulation set considers the attacked process. The attack is classified based on its detectability. Applying Proposition 10, the attack is detectable because $\mathcal{R}_k^{\eta_a}(\mathcal{R}_k^{\xi_a})$ and $\mathcal{R}_k^{\eta}(\mathcal{R}_k^{\xi})$ do not intersect at $k = 0$ and $k = 1$, as depicted in Fig. 5.5a (the sets at $k = 0$ do not intersect, even though they appear to intersect at their boundary). Several closed-loop simulations are performed to verify that the attack is detected in all simulations. The detection scheme raises an alarm in all simulations at $k = 0$ and $k = 1$. For some simulations, an alarm is raised over subsequent time steps, but the attack is no longer detected over time once the augmented state converges to the minimum invariant set, i.e., the alarm goes away over time. This behavior occurs because $\mathcal{R}_\infty^{\eta} \subset \mathcal{R}_\infty^{\eta_a}$ (albeit this is difficult to see from Fig. 5.5a), but the non-intersecting area between the two sets is small. Figure 5.5b illustrates the values of the monitoring variable over one simulation of the attacked process. Over this simulation, the attack is detected at all $k \in [0, 4]$. For $k \in [5, 5000]$, the monitoring variable evolves within the reachable sets of the attack-free process, and no alarms are raised.

### 5.3.2 Factors that Influence the Detectability of a Multiplicative False Data Injection Attack

In this case study, a multiplicative attack that alters the data communicated over the sensor-controller and controller-actuator links with pre-multiplication factors $\Lambda^y = 1.1$, $\Lambda^u = 0.9$, and biases $\delta_k^y = \delta_k^u = 0$ for all $k \in \mathbb{Z}^+$ is considered. The impact of the initial set and controller-observer gains on the detectability of this attack is explored. The process evolving from two different initial sets is considered to explore the impact of the initial set on the attack detectability. The first initial set is the set that is the attack-free minimum invariant set shifted by $\xi' = [10 \ \ -5]^T$, and the second is the set that is the the attack-free minimum invariant set by $\xi' = [50 \ \ -20]^T$. The process is operated with controller-observer gains of $K = 1$ and $L = 0.9$. Figure 5.6a and Fig. 5.6b illustrate the reachable sets of the monitoring variable for the attack-free and the attacked processes for a few time steps starting from the first and second initial set, respectively. For the given controller-observer gains, the attack is potentially detectable or detectable with respect to the first initial set because the two sets intersect for all time, and the attacked reachable set is not a subset of the attack-free reachable set (Fig. 5.6a). The attack, however, is detectable with respect to the second initial set because the reachable sets do not intersect at $k = 1$ and $k = 2$ (Fig. 5.6b).

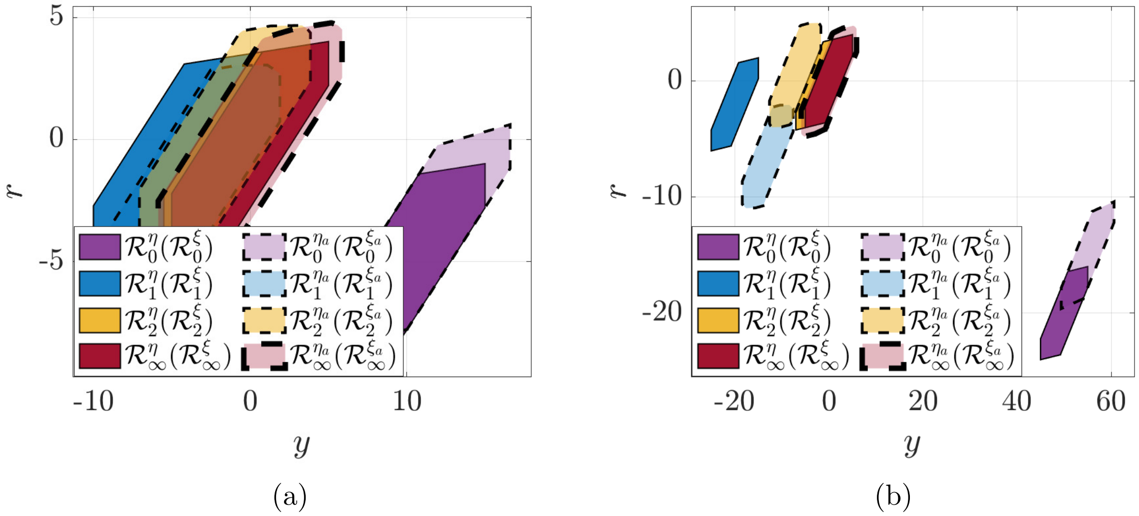
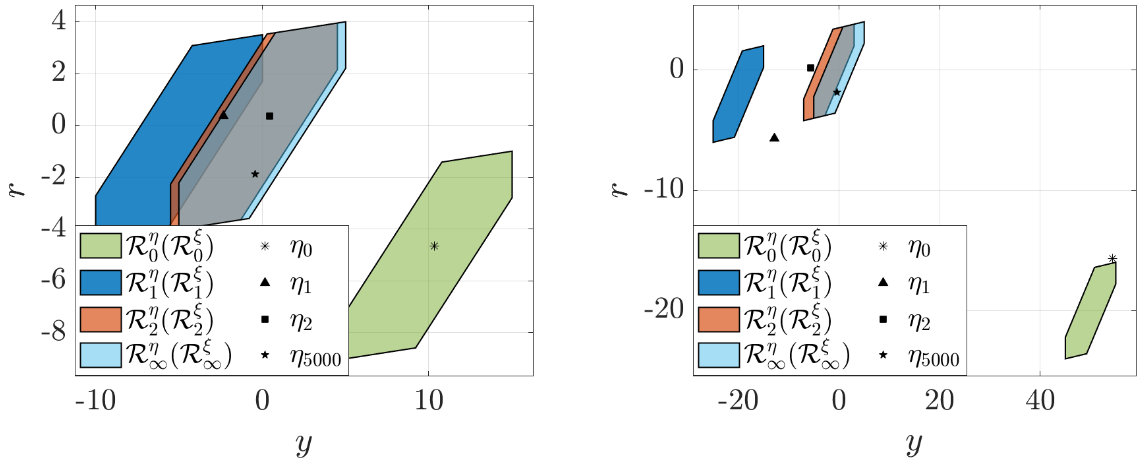


Fig. 5.6: The reachable sets over a few time steps of the attack-free process and process under a multiplicative FDIA. Two initial sets are considered: (a) an initial set that is the attack-free minimum invariant set shifted by $\xi' = [10 \;\; -5]^T$ and (b) an initial set that is the attack-free minimum invariant set shifted by $\xi' = [50 \;\; -20]^T$.

To further confirm these findings, two sets of simulations of the attacked process are performed. In the first simulation set, the process evolving from the initial condition $\xi_0 = [10 \;\; -5]^T$, contained in the first initial set, is considered. In the second simulation set, the process evolving from the initial condition $\xi_0 = [50 \;\; -20]^T$, contained in the second initial set, is considered. Each simulation set consists of 1000 simulations of the attacked process. In the first simulation set, the attack is detected over 474 of the 1000 simulations. For the simulations where the attack is detected, the first detection time ranged from $k = 1$ to $k = 4970$, indicating a range of detection times. The monitoring variable values and reachable sets used in the detection scheme are shown in Fig. 5.7a for one simulation. Over this simulation, the monitoring variable values over the time steps shown are contained within the reachable sets used in the detection scheme, and the attack is not detected. In the second simulation set, the attack is detected in all simulations at $k = 1$ and $k = 2$. The monitoring variable values and reachable sets used in the detection scheme over a few time steps are shown in Fig. 5.7b for one simulation where the attack is detected at $k = 0$, $k = 1$, and $k = 2$. These results demonstrate the dependence of the attack detectability on the initial set.

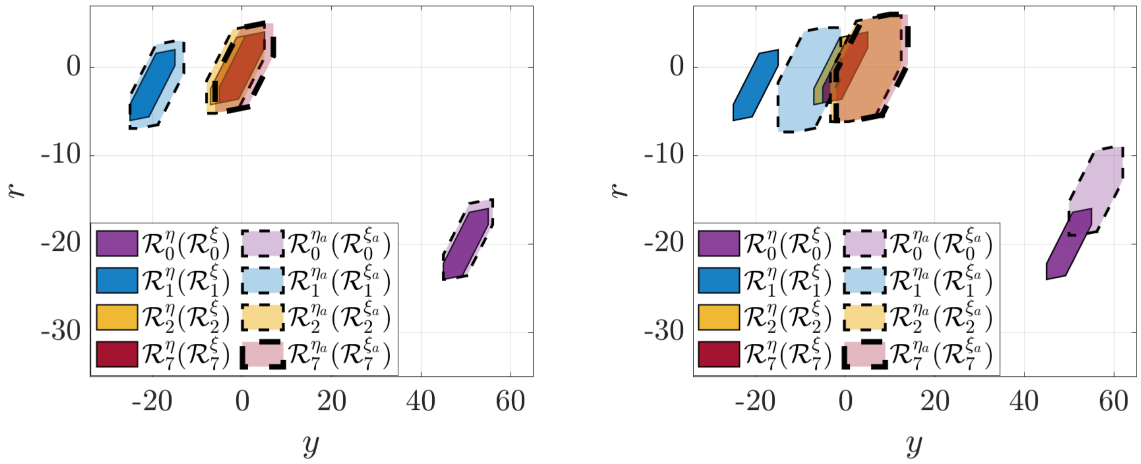


(a)

(b)

Fig. 5.7: The monitoring variable values and reachable sets used in the detection scheme over a few time steps for the process under a multiplicative FDIA. Two initial sets are considered: (a) an initial set that is the attack-free minimum invariant set by $\xi' = [10 \ -5]^T$ and (b) an initial set that is the attack-free minimum invariant set shifted by $\xi' = [50 \ -20]^T$.

The impact of the controller-observer gains on attack detectability is also analyzed by considering process operation for two choices of the controller-observer gains: $(K, L) = (1, 0.9)$ and $(K, L) = (0.2, 1.5)$. The process evolving from an initial set that is the attack-free minimum invariant set shifted by $\xi' = [50 \ -20]^T$ is considered. As described above, the attack is detectable when the process is operated with $(K, L) = (1, 0.9)$. Applying the attack classification scheme, the attack is detectable or potentially detectable when the process is operated with $(K, L) = (0.2, 1.5)$. An additional 1000 simulations of the attacked process under the second choice of gains are performed. The attack is detected in 638 of the 1000 simulations. However, the attack is detected in all 1000 simulations when the process is operated with the first choice of gains. These results indicate that the choice of controller-observer gains can also influence the ability to detect attacks.

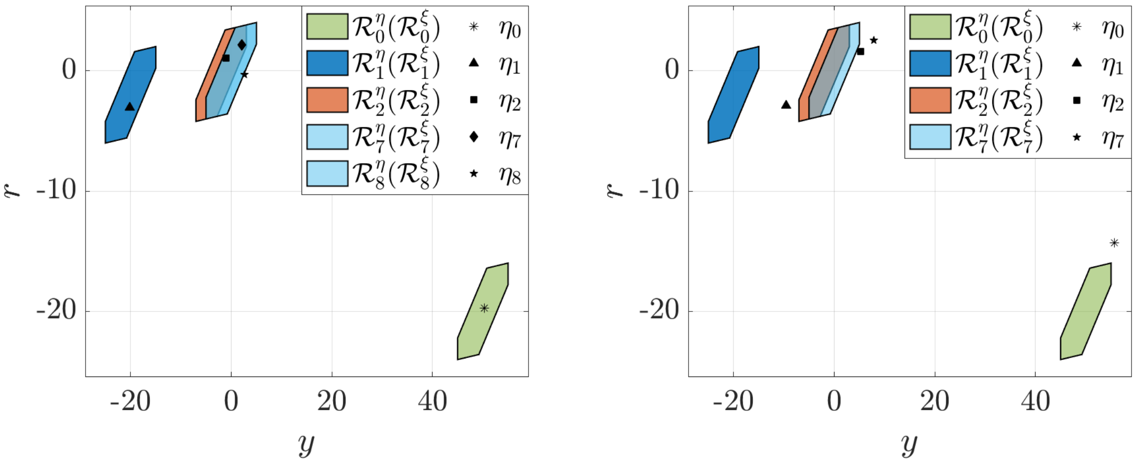


(a) (b)

Fig. 5.8: The reachable sets over a few time steps for the attack-free process with respect to the reachable sets for the process under (a) the first additive attack and (b) the second additive attack.

### 5.3.3 Factors Influencing the Detectability of an Additive FDIA

In this case study, additive FDIAs (i.e., attacks with $\Lambda^y = 1$ and $\Lambda^u = 1$) are considered. First, the detectability of two additive attacks with respect to the reachable set-based detection scheme is analyzed. Next, the impact of the initial set on attack detectability is analyzed. Finally, the influence of the controller-observer gains on the detectability of an additive attack is analyzed.



(a)

(b)

Fig. 5.9: The monitoring variable values and reachable sets used in the detection scheme over a few time steps for the attacked process. The monitoring variable values shown are observed over one simulation of the process under: (a) the first additive attack and (b) the second additive attack.

The detectability of two additive FDIAs that alter the variable values over the sensor-controller and controller-actuator links is analyzed. Both attacks involve randomly varying $\delta_k^u$ and $\delta_k^y$ where both values are drawn from a uniform distribution at every time step. For the first attack, both numbers are drawn from the interval $[0, 1]$, and for the second attack, both numbers are drawn from the interval $[5, 7]$. The process is operated with $K = 1$ and $L = 0.9$ and an initial set that is the attack-free minimum invariant set shifted by $\xi' = [50\ -20]^T$. The reachable sets of the attack-free process converge to its minimum invariant set at the time step $k = 7$. Figure 5.8a and Fig. 5.8b illustrate the reachable sets of the monitoring variable for the attack-free process with respect to the reachable sets of the process under the first attack and the second attack, respectively. As illustrated in Fig. 5.8a, the first attack is either detectable or potentially detectable because the attacked reachable sets intersect, but are not contained within the attack-free reachable sets at all time steps. However, the second attack is detectable because the attacked and the attack-free reachable sets do not intersect at $k = 1$ (Fig. 5.8b).

To investigate attack detectability further, two sets of closed-loop simulations of the process under an attack are performed. In the first simulation set, the process under the first attack is considered. In the second simulation set, the process under the second attack is considered. Each simulation set consists of 1000 simulations of the process. All simulations are initialized at $\xi_0 = [50 \;\; -20]^T$, which is within the initial set. The first attack is detected in all simulations, with the detection time ranging from $k = 2$ to $k = 1402$. Figure 5.9a illustrates the attack-free reachable sets for a few time steps, and the monitoring variable values over one simulation. Over this simulation, the monitoring variable values are contained within the reachable sets from $k = 0$ to $k = 7$. The attack is detected at time step $k = 8$. On the other hand, the second attack is detected at time step $k = 1$ in all simulations. Figure 5.9a illustrates the attack-free reachable sets and the monitoring variable values over one simulation. Over this simulation, the attack is detected at all time steps shown (i.e, from $k = 0$ to $k = 7$). The results demonstrate that an additive attack of this nature, where the attack bias is treated as a random number within compact interval, may be detectable or potentially detectable.

Next, the impact of the initial set on the detectability of an additive attack is analyzed by considering the process operated with $K = 1$ and $L = 0.9$. The process evolving from three different initial sets is considered: first from an initial set that is the attack-free minimum invariant set shifted by $\xi'_1 = [10 \;\; -10]^T$, second from an initial set that is the attack-free minimum invariant set shifted by $\xi'_2 = [100 \;\; -50]^T$, and third from an initial set that is the attack-free minimum invariant set shifted by $\xi'_3 = [50 \;\; -20]^T$. For the process evolving from each initial set, the detectability of the two additive attacks considered previously is analyzed. For the process evolving from all initial sets considered, the first attack where the random attack biases are bounded in $[0, 1]$ is found to be either potentially detectable or detectable. For the process evolving from all three initial sets considered, the second attack where the random attack biases are bounded in $[5, 7]$ is detectable, because the reachable sets of the attacked and the attack-free process do not intersect at the time step $k = 1$. The results demonstrate that for the process evolving from any of the initial sets considered, the detectability of the two additive attacks is consistent.

Finally, the impact of the controller-observer gains on the detectability of an attack is explored by considering the process evolving from the initial set that is the attack-free minimum invariant set shifted by $\xi' = [50 \quad -20]^T$ and operated with two controller-observer gains: first with $K = 0.2$ and $L = 1.5$ and second with $K = 1$ and $L = 0.9$. For the process operated with each choice of controller-observer gains, the detectability of the additive attack where the attack biases are bounded in $[5, 7]$ is analyzed. For the process operated with $K = 0.2$ and $L = 1.5$, the attack may either be detectable or potentially detectable. However, the attack on the process operated with $K = 1$ and $L = 0.9$ is detectable with the reachable sets of the attacked and the attack-free process having zero intersection at time step $k = 1$. The results demonstrate that the controller-observer gains influence the detectability of an additive FDIA.

## 5.4 Numerical Results: Chemical Process Example

In this section, the proposed reachable set-based detection scheme, and the detectability-based classification of attacks, are applied to a chemical process example during transient operation. All polytope computations are performed using the MPT 3.0 toolbox ([82]). A chemical process example consisting of a well-mixed continuously stirred-tank reactor (CSTR) is considered. A second-order exothermic reaction $A \rightarrow B$ occurs in the CSTR. Under standard modeling assumptions, the process dynamics are described by its mass and energy balances:

$$\begin{aligned}\frac{dC_A}{dt} &= \frac{F}{V}(C_{A0} + \Delta C_{A0} - C_A) - k_0 e^{\frac{-E}{RT}} C_A^2 \\ \frac{dT}{dt} &= \frac{F}{V}(T_0 + \Delta T_0 - T) - \frac{\Delta H k_0}{\rho C_p} e^{\frac{-E}{RT}} C_A^2 + \frac{Q}{\rho C_p V}\end{aligned} \tag{5.18}$$

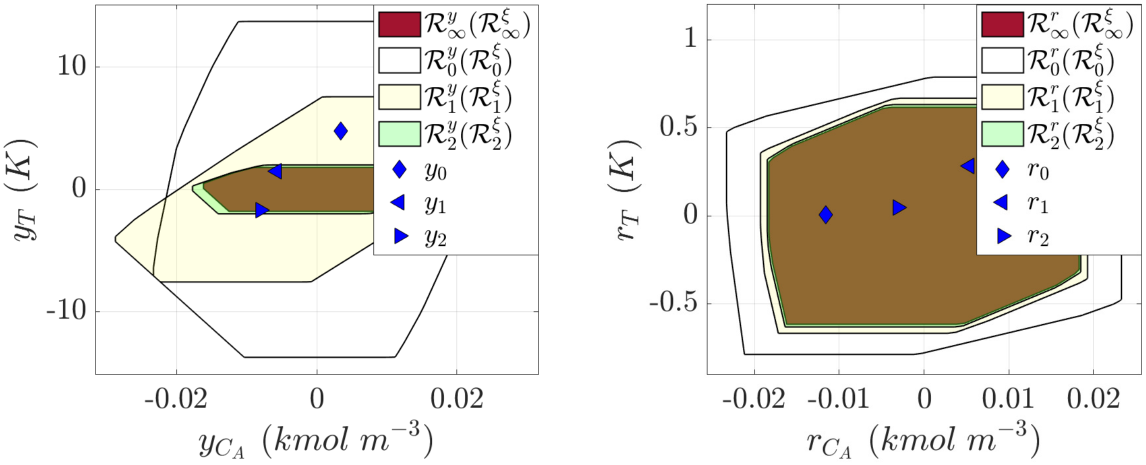
where $C_{A0}$ and $T_0$ are the reactant feed concentration and feed temperature, respectively; and $C_A$ and $T$ are the concentration of the reactant in the reactor and the temperature of the reactor, respectively. The manipulated input is the heat supplied to or removed from the reactor $Q$. The process is subject to bounded disturbances modeled as the variation in the concentration of the reactant $A$ in the feed $\Delta C_{A0}$, and the variation of the temperature of the feed to the reactor $\Delta T_0$. The measured variables available to the controller are the concentration of the reactant $C_A$ and the temperature of the reactant $T$. The process is subject to bounded measurement noise acting on all sensors. The

Process parameters of the CSTR

| | |
|---|---|
| Volumetric flow rate ($F$) | $5.0\,\mathrm{m^3\,h^{-1}}$ |
| Reactor volume ($V$) | $1.0\,\mathrm{m^3}$ |
| Feed concentration of $A$ ($C_{A0}$) | $4.0\,\mathrm{kmol\,m^{-3}}$ |
| Activation energy ($E$) | $5.0 \times 10^4\,\mathrm{kJ\,kmol^{-1}}$ |
| Pre-exponential factor ($k_0$) | $8.46 \times 10^6\,\mathrm{m^3\,h^{-1}\,kmol^{-1}}$ |
| Gas constant ($R$) | $8.314\,\mathrm{kJ\,kmol^{-1}\,K}$ |
| Feed temperature ($T_0$) | $300\,\mathrm{K}$ |
| Density of reactor liquid hold-up ($\rho$) | $1000\,\mathrm{kg\,m^{-3}}$ |
| Heat of reaction ($\Delta H$) | $-1.15 \times 10^4\,\mathrm{kJ\,kmol^{-1}}$ |
| Heat capacity ($C_p$) | $0.231\,\mathrm{kJ\,kg\,K^{-1}}$ |
| Steady-state heat rate added/removed from the reactor ($Q_s$) | $0\,\mathrm{kJ\,h^{-1}}$ |
| Steady-state reactant concentration ($C_{As}$) | $1.22\,\mathrm{kmol\,m^3}$ |
| Steady-state temperature ($T_s$) | $438.2\,\mathrm{K}$ |

process disturbances are bounded such that $|\Delta C_{A0}| \leq 0.01\,\mathrm{kmol\,m^{-3}}$ and $|\Delta T_0| \leq 0.2\,\mathrm{K}$. Similarly, the measurement noise acting on the concentration sensor ($v_1$) is bounded as $|v_1| \leq 0.01\,\mathrm{kmol\,m^{-3}}$, and the measurement noise on the temperature sensor ($v_2$) is bounded as $|v_2| \leq 0.2\,\mathrm{K}$. The definitions and values of the other process parameters are listed in Table 2.1, and are reproduced in this chapter to make it self-contained.

The control objective is to stabilize the closed-loop process at its open-loop stable steady state given by $C_{As} = 1.22\,\mathrm{kmol\,m^{-3}}$, $T_s = 438\,\mathrm{K}$, and $Q_s = 0\,\mathrm{kW}$. A continuous-time linear time-invariant state-space model is obtained via linearization around the desired operating steady-state of the CSTR, and defining the deviation variables $x_1 = C_A - C_{As}$, $x_2 = T - T_s$, and $u = Q - Q_s$. Using a sampling interval of $\Delta = 1 \times 10^{-2}\,\mathrm{h}$, a discrete-time state-space model of the form in Eq. 5.1 is obtained. A monitoring variable that is the concatenation of the measured output and the residual vectors is considered, i.e., $\eta_k := [y_k^T \; r_k^T]^T$. In the case studies that follow, the process is simulated using its continuous-time nonlinear model in Eq. 5.18 with the control input applied in a sample-and-hold fashion. Euler's method with an integration step size of $1 \times 10^{-4}$ h is used to integrate the ordinary differential equations. Two case studies are performed. In the first case study, the reachable set-based detection scheme is applied to monitor the CSTR during a transient phase induced by switching the controller-observer gains during operation. In the second case study, the detectability of a simultaneous additive and multiplicative FDIA is analyzed using the reachable set-based attack detectability classification scheme. For both case studies, the linearized process model is used to design the control law and compute the reachable sets. However, the CSTR is simulated using its nonlinear model. Therefore, the case studies presented in this section consider the application of the classification of attacks and the detection scheme to a nonlinear process.



(a)

(b)

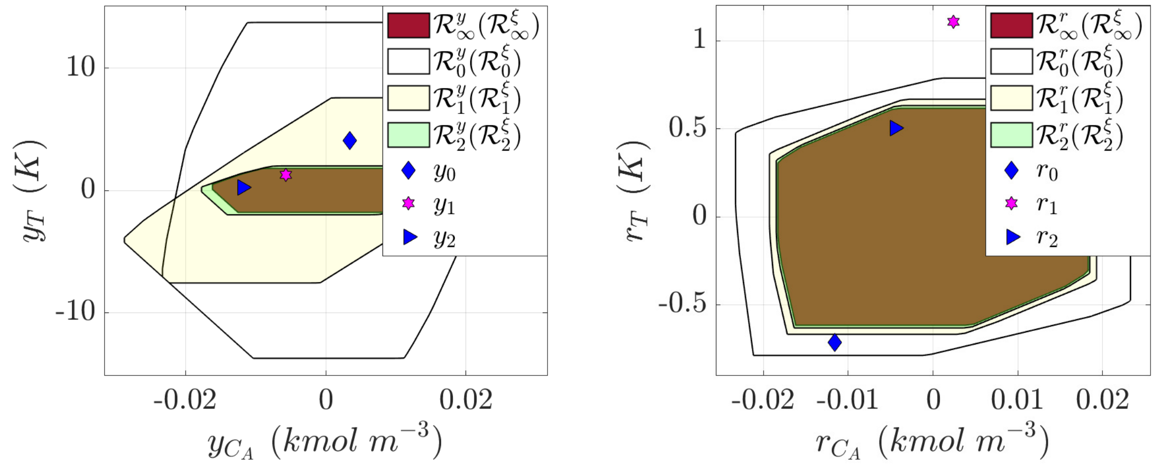
Fig. 5.10: The monitoring variable values, including (a) the output values and (b) the residual values, and reachable sets used in the detection scheme over a few time steps for the attack-free process. In this case, there are no false alarms. The brown central region represents the intersection of all reachable sets shown.

### 5.4.1 Application of the Reachable Set-Based Detection Scheme

In a prior work ([48]), controller-observer gain switching between $(K_i, L_i)$, with controller poles at $[-0.2\ -0.3]$ and observer poles at $[-0.2\ -0.3]$, to $(K_f, L_f)$, with controller poles at $[0.2\ -0.1]$ and observer poles at $[0.2\ 0.3]$, was considered as a way to enhance attack detection capabilities of a detection scheme monitoring the process. In this case study, gain switching occurs on the process operating initially with its states bounded in the minimum invariant set of the attack-free process under $(K_i, L_i)$. The controller switch may induce a transient operation, so the reachable set-based detection scheme is applied to monitor the process. The forward reachable sets of the attack-free process from the minimum invariant set of the attack-free process under $(K_i, L_i)$, which is taken to be the initial set $\mathcal{R}_0^\xi$, are computed, and the reachable sets converge to the minimum invariant set of the attack-free process under $(K_f, L_f)$, which is denoted by $\mathcal{R}_\infty^\xi$. To design the reachable set-based detection scheme, the reachable sets of the monitoring variable for the attack-free process are computed from the initial set $\mathcal{R}_0^\eta(\mathcal{R}_0^\xi)$ to its terminal set $\mathcal{R}_\infty^\eta(\mathcal{R}_\infty^\xi)$. Two sets of simulations are considered. The first set considers the attack-free process, and the second set considers the process under a multiplicative sensor-controller link attack of magnitude $\Lambda^y = \text{diag}(1, 0.85)$. Each simulation set consists of 1000 simulations of the process, and each simulation has a total length of 5 h, spanning 500 time steps. All simulations are initialized with $\xi_0 = [0.005\ 5\ \ -0.01\ 0.2]^T \in \mathcal{R}_0^\xi \setminus \mathcal{R}_\infty^\xi$.

No attacks are detected using the reachable set-based detection scheme when monitoring the attack-free process. Figure 5.10a and Fig. 5.10b illustrate the output and the residual values over a few time steps for one of the simulations of the attack-free process. The monitoring variable values are contained within their corresponding reachable sets for all time. At 0.02 h ($k = 2$), the monitoring variable values converge to the terminal set for the attack-free process, where they remain. As a result, the reachable set-based detection scheme generates an output of 0 for all time steps in the simulation.

Considering the process under a multiplicative attack, the attack is detected in 854 out of the 1000 simulations. The detection times ranged from 0.01 h ($k = 1$) to 4.59 h ($k = 459$) for the simulations where the attack is detected. Figure 5.11a and Fig. 5.11b illustrate the output and residual values over a few time steps over a simulation of the attacked process. From Fig. 5.11b, the attack is detected at 0.01 h ($k = 1$). These simulations demonstrate that the reachable set-based detection scheme can monitor the nonlinear process during transient operation without raising false alarms for the attack-free process, and can successfully detect attacks on a nonlinear process.



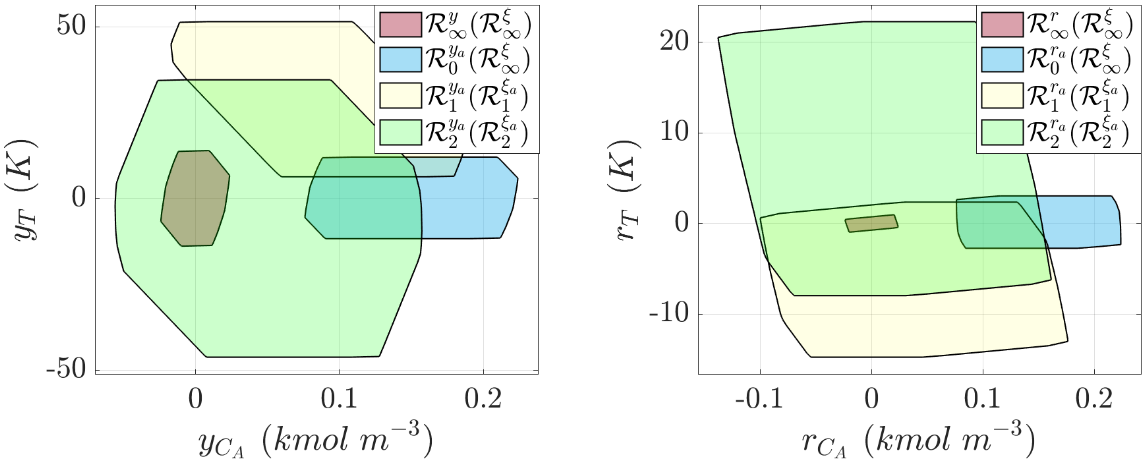
(a) (b)

Fig. 5.11: The monitoring variable values, including (a) the output values and (b) the residual values, and reachable sets used in the detection scheme over a few time steps for the attacked process. In this case, the attack is detected at $k = 1$. The brown central region represents the intersection of all reachable sets shown.

**Remark 5.4.1.** *In prior work ([48]), a controller-observer gain switch was utilized to enable attack detection on the nonlinear CSTR process monitored by a terminal set-based detection scheme. However, the attack detection method presented previously has a non-zero false alarm rate because the terminal set-based detection scheme is not designed to account for transient operation. Based on the results in this section, the reachable set-based detection scheme may be used to eliminate false alarms in the controller-observer gain switching-based attack detection method.*

### 5.4.2 Application of Detectability-Based Classification of an Attack

In this case study, the ability to classify attacks using the reachability analysis is demonstrated for the nonlinear CSTR. Specifically, the detectability of a simultaneous multiplicative and additive FDIA that alters the data over both the sensor-controller and controller-actuator links is analyzed. The attack parameters are $\Lambda^y = \mathrm{diag}(1, 0.85)$, $\Lambda^u = 0.9$, $\delta_k^{y_{C_A}} \in [0.1, 0.2]$ kmol m$^{-3}$, and $\delta_k^{y_T} \in [0.1, 0.2]$ K. The parameters $\delta_k^{y_{C_A}}$ and $\delta_k^{y_T}$ are the attack biases added to the concentration and temperature measurements, respectively, and are modeled as random variables drawn from a uniform distribution. The process is operated with controller-observer gains selected via pole placement using the linearized process model and by placing the poles at $[-0.2 \ -0.3]$ to determine $K$ and $[-0.2 \ -0.3]$ to determine $L$. For the attack-free process, the minimum invariant set of the closed-loop system is the initial set, so the terminal set of the monitoring variable is the $k$-step forward reachable set for all time steps $k \in \mathbb{Z}^+$.



(a)

(b)

Fig. 5.12: The reachable sets for (a) the measured output and (b) the residual for the CSTR under an attack.

The closed-loop process under the FDIA is unstable ($\max_i|\lambda_i(A^{\xi_a})| = 1.1371 > 1$). The reachable sets of the attacked process are compared to the terminal set of the attack-free process to classify the attack. Fig 5.12a illustrates the reachable sets of the measured output for the attacked process for a few time steps and the measured output terminal set for the attack-free process. Figure 5.12b illustrates the reachable sets of the residual for the attacked process and the residual terminal set for the attack-free process. As illustrated, the attack is detectable with respect to the initial set because the reachable set of the attacked process and the terminal set of the attack-free process do not intersect at time step $k = 0$.

Two simulation sets are performed to confirm that the reachability analysis correctly classified the attack. The attack-free process is considered in the first set, and the attacked process is considered in the second set. Each set consists of 1000 simulations of the process, and each simulation simulates the CSTR over a 5 h period (total of 500 time steps). All simulations are initialized with the augmented state at the origin. The detection scheme in Eq. 5.14 is designed to monitor the process with respect to the reachable sets, which are the terminal set of the attack-free process ($\mathcal{R}_k^{\eta}(\mathcal{R}_k^{\xi}) = \mathcal{R}_{\infty}^{\eta}(\mathcal{R}_{\infty}^{\xi})$ for all time steps $k \in \mathbb{Z}^+$).

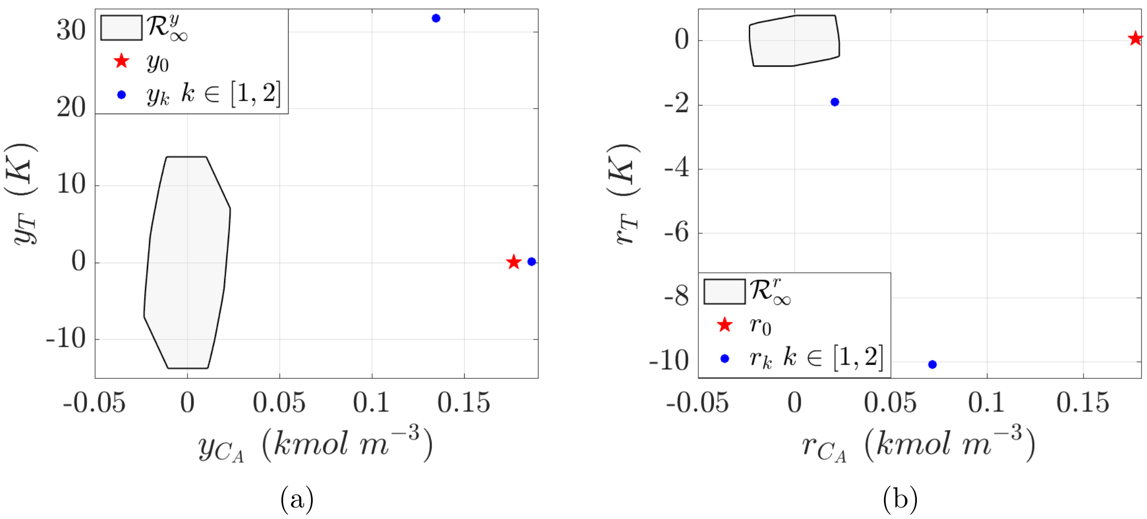


Fig. 5.13: The values of (a) the measured output and (b) the residual and their corresponding terminal sets over one simulation of the CSTR process under an attack.

For the attack-free simulations, the detection scheme does not raise any alarms. For the simulations of the attacked process, the attack is detected at $k = 0$ in all simulations, as expected from the reachability analysis. The measured output and the residual (monitoring variable) values for the attacked process over one simulation are shown in Fig. 5.13a and Fig. 5.13b, respectively. Over this simulation, the monitoring variable values evolve outside the terminal set of the attack-free process but stay within the attacked process reachable sets. The attack is detected at the first three time steps. The results demonstrate that the detectability classification based on reachable sets can be applied to classify attacks for the nonlinear CSTR.

**Remark 5.4.2.** *For the attack-free process, the terminal set of the monitoring variable is the forward reachable set for all time steps $k \in \mathbb{Z}^+$. Therefore, the terminal set-based detection scheme is a special case of the reachable set-based detection scheme with*

$\mathcal{R}_k^{\eta}(\mathcal{R}_k^{\xi}) = \mathcal{R}_{\infty}^{\eta}(\mathcal{R}_{\infty}^{\xi})$ *(for all $k \in \mathbb{Z}^+$) in Eq. 5.14, and the reachable set-based classification of attacks presented in Section 5.2 can be used to analyze attack detectability for a process monitored by the terminal set-based detection scheme.*

## 5.5 Conclusions

A reachable set-based detection scheme was proposed to monitor dynamic processes under false data injection attacks targeting the sensor-controller and controller-actuator communication links. A rigorous characterization of reachable set-based conditions that result in an attack being undetectable or detectable with respect to the proposed detection scheme was presented. An approach for classifying attacks based on their detectability with respect to the reachable set-based detection scheme was presented. The proposed detection scheme was applied to two illustrative examples. The detectability of various attacks was analyzed, and the applicability of the detection scheme and classification method to monitor and classify attacks on a nonlinear chemical process was demonstrated.

# Chapter 6

# Detection of Multiplicative False Data Injection Cyberattacks on Process Control Systems via Randomized Control Mode Switching

In this chapter, a reachable set-based detection scheme [49] that monitors a process under transient operation for attacks by tracking the values of the monitoring variable with respect to its attack-free reachable sets to eliminate false alarms from a switch is used. To enable the detection of a range of multiplicative FDI attacks, a randomized switching-enabled cyberattack detection method under which the control mode is switched at randomly chosen switching instances is proposed. Randomization to enhance the cybersecurity of a control system has received some attention in the literature [93–96]. However, randomization to enhance the cybersecurity of the detection scheme has not been explored. Under the proposed randomized detection method, the control mode switching instances are randomly chosen to confound an attacker aiming to learn the switching schedule. Without complete knowledge of the switching schedule, an attacker may not be able to design a "smart" attack that is capable of evading detection, thereby, the cybersecurity of the detection method is enhanced. Some practical implementation issues are discussed, and two variations of the algorithms for the active detection method are presented. The application of the active detection method in enabling attack detection while guaranteeing a zero false alarm rate is demonstrated using simulations of two illustrative processes. Finally, using simulations of the second illustrative example, where the detection method is applied to a chemical process operated at its steady-state, the detection of a smart attack is demonstrated. The results demonstrate that the randomized switching-enabled attack detection method may be preferred to a method using scheduled control mode switches.

## 6.1 Preliminaries

### 6.1.1 Notation and Definitions

$\mathbb{Z}^+$ is the set of non-negative integers. $\mathbb{R}^n$ is the n-dimensional Euclidean space. Given a vector $x = [x_1 \ x_2 \ \ldots \ x_n]^T \in \mathbb{R}^n$, $\|x\| = \sqrt{\sum_{i=1}^{n}(x_i)^2}$ is its Euclidean norm and $\|x\|_\infty = \max_{i=1}^{n}|x_i|$ is its $\infty$-norm. For a square matrix $A \in \mathbb{R}^{n\times n}$, its spectral radius is defined as $\rho(A) = \max\{|\lambda_1|, |\lambda_2|, \ldots, |\lambda_n|\}$, where $\lambda_i$ is the $i^{\text{th}}$ eigenvalue of the matrix $A$. The Minkowski sum of two sets $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^n$ is defined as $\mathcal{X} \oplus \mathcal{Y} = \{x + y \mid x \in \mathcal{X}, y \in \mathcal{Y}\}$. Given a matrix $A \in \mathbb{R}^{m\times n}$ and a set $\mathcal{X} \subset \mathbb{R}^n$, $A\mathcal{X} = \{Ax \mid x \in \mathcal{X}\}$ is the linear map of the set $\mathcal{X}$. Given a matrix $A \in \mathbb{R}^{m\times n}$ and a set $\mathcal{X} \subseteq \mathbb{R}^n$, $\bigoplus_{i=0}^{t} A^i\mathcal{X}$ represents the series $\mathcal{X} \oplus A\mathcal{X} \oplus \cdots \oplus A^t\mathcal{X}$. Given a vector $x \in \mathbb{R}^n$, an $n$-dimensional polytope is a bounded region in the Euclidean space that satisfies the linear matrix inequalities $Ax \leq b$, where $A \in \mathbb{R}^{m\times n}$ is a matrix and $b \in \mathbb{R}^{m\times 1}$ is a vector. A zonotope is a convex polytope that is symmetric about its center and may be formally defined as the Minkowski sum of a finite set of line segments [97].

### 6.1.2 Class of Processes

In this chapter, processes that are modeled by discrete-time linear time-invariant systems of the following form are considered:

$$x_{t+1} = A^x x_t + B^u u_t + B^w w_t \tag{6.1a}$$

$$y_t = C^x x_t + v_t \tag{6.1b}$$

where $x_t \in \mathbb{R}^{n_x}$ and $u_t \in \mathbb{R}^{n_u}$ for all $t \in \mathbb{Z}^+$ are the vectors representing the process states and manipulated inputs, respectively, and the measured output is $y_t \in \mathbb{R}^{n_y}$ with $n_y \leq n_x$. The process is subject to bounded process disturbances and measurement noise, where $w_t \in \mathcal{W} \subset \mathbb{R}^{n_x}$ and $v_t \in \mathcal{V} \subset \mathbb{R}^{n_y}$ are the vectors representing the bounded process disturbances and measurement noise. The compact sets $\mathcal{W}$ and $\mathcal{V}$ are assumed to be known polytopes. $A^x \in \mathbb{R}^{n_x \times n_x}$, $B^u \in \mathbb{R}^{n_x \times n_u}$, $B^w \in \mathbb{R}^{n_x \times n_w}$, and $C^x \in \mathbb{R}^{n_y \times n_x}$ are matrices.

To estimate the process states, a Luenberger observer is utilized:

$$\hat{x}_{t+1} = A^x \hat{x}_t + B^u u_t + L(y_t - \hat{y}_t) \tag{6.2a}$$

$$\hat{y}_t = C^x \hat{x}_t \tag{6.2b}$$

where $\hat{x}_t \in \mathbb{R}^{n_x}$ and $\hat{y}_t \in \mathbb{R}^{n_y}$ for all $t \in \mathbb{Z}^+$ are the estimates of the process states and the measured outputs generated by the observer. The observer gain $L \in \mathbb{R}^{n_x \times n_y}$ is chosen such that all eigenvalues of the matrix $A - LC$ are strictly within the unit circle. The estimates of the process states generated by the observer are utilized to compute the control action as follows:

$$u_t = -K\hat{x}_t \tag{6.3}$$

where $K \in \mathbb{R}^{n_x \times n_u}$ is the gain of the feedback controller, which is chosen to ensure that all eigenvalues of the matrix $A - BK$ are contained strictly inside the unit circle.

The discrepancy between the process states and their estimates generated by the observer is defined as the estimation error ($e \coloneqq x - \hat{x}$) and has the dynamics:

$$e_{t+1} = (A^x - LC^x)e_t + B^w w_t - Lv_t \tag{6.4}$$

The dynamics of the overall closed-loop process are collectively described by the evolution of the process states and the estimation errors, which are analyzed using an augmented state $\xi = [x^T \ e^T]^T$, which consists of the process states and estimation errors. The dynamics of the closed-loop augmented state are governed by:

$$\xi_{t+1} = \underbrace{\begin{bmatrix} A^x - B^u K & B^u K \\ 0 & A^x - LC^x \end{bmatrix}}_{A^\xi(K,L)} \xi_t + \underbrace{\begin{bmatrix} B^w & 0 \\ B^w & -L \end{bmatrix}}_{B^d(L)} d_t \tag{6.5}$$

where $d_t := [w_t^T \;\; v_t^T]^T$ is an augmented disturbance vector, which consists of the process disturbances and measurement noise and is bounded within the compact set $\mathcal{D} := \{[w^T \;\; v^T]^T \mid w \in \mathcal{W}, v \in \mathcal{V}\}$. Without loss of generality, the origin is assumed to be the steady state of the process.

In the remainder of the paper, the term "attack-free closed-loop process with $(K, L)$ and an initial set of states $\mathcal{R}_0^\xi$" is used to refer to the process with augmented state dynamics, as in Eq. 6.5. In this chapter, process transients are considered, during which the states of the closed-loop process in Eq. 6.5 evolve from a compact initial set of states ($\xi_0 \in \mathcal{R}_0^\xi \subset \mathbb{R}^{2n_x}$). For the closed-loop process with $(K, L)$, its $t$-step reachable set is defined as the set of all states that can be reached in $t$ time steps from the initial set of states $\mathcal{R}_0^\xi$ and under all admissible disturbances [98]. The $t$-step reachable set of the closed-loop process may be expressed as:

$$\mathcal{R}_t^\xi(K, L) = A^\xi(K, L)\mathcal{R}_{t-1}^\xi(K, L) \oplus B^d(L)\mathcal{D}, \text{ for } t > 0 \tag{6.6}$$

From Eq. 6.6, the $t$-step reachable set of the process is dependent on the initial set of states $\mathcal{R}_0^\xi$, the control parameters $(K, L)$, the time step $t \in \mathbb{Z}^+$, and the set of bounded disturbances $\mathcal{D}$. For a concise representation, in this section, the highlight is on the dependence of the $t$-step reachable set on the control parameters only. When the closed-loop process with $(K, L)$ is stable in the sense that $\rho(A^\xi(K, L)) < 1$, its augmented state is ultimately bounded within the minimum invariant set of the process, which is the limit set of all trajectories of the process [79]. The minimum invariant set may be used to analyze the behavior of the closed-loop process under steady-state operation and can be expressed as the infinite Minkowski sum [77]:

$$\mathcal{R}_\infty^\xi(K, L) = \bigoplus_{i=0}^{\infty} A^\xi(K, L)^i B^d(L)\mathcal{D} \tag{6.7}$$

### 6.1.3 Class of Multiplicative False Data Injection Attacks

In this chapter, the detection of multiplicative false data injection (FDI) attacks that alter the data communicated over the sensor-controller and controller-actuator links of the PCS network simultaneously is considered. The falsified data are represented as follows:

$$y_t^a = \Lambda^y y_t \tag{6.8a}$$

$$u_t^a = \Lambda^u u_t \tag{6.8b}$$

where $\beta_t^a \in \mathbb{R}^{n_\beta}$ is the value of the variable $\beta \in \mathbb{R}^{n_\beta}$ that is altered by the attack, $\Lambda^\beta \neq I$ is the pre-multiplicative factor that alters the value of $\beta$ in the presence of an attack. If $\beta = y$, the attack alters the value of the measured output communicated to the controller, and if $\beta = u$, the attack alters the value of the manipulated input communicated to the control actuators. In the presence of an attack, $u^a$ is the implemented control action that is used by the observer in Eq. 6.2a to generate estimates of process states. An attack alters the dynamics of the closed-loop process in Eq. 6.5 as follows:

$$\xi_{t+1} = \underbrace{\begin{bmatrix} A^x - B^u\Lambda^u K & B^u\Lambda^u K \\ L(I-\Lambda^y)C^x & A^x - LC^x \end{bmatrix}}_{A^{\xi_a}(K,L)} \xi_t + \underbrace{\begin{bmatrix} B^w & 0 \\ B^w & -L\Lambda^y \end{bmatrix}}_{B^{d_a}(L)} d_t \tag{6.9}$$

In the remainder of the paper, the term "attacked closed-loop process with $(K, L)$ and an initial set of states $\mathcal{R}_0^\xi$" is used to refer to the process under an attack with augmented state dynamics, as in Eq. 6.9. The $t$-step reachable set of the attacked process with $(K, L)$ and an initial set of states $\mathcal{R}_0^\xi$ may be expressed as:

$$\mathcal{R}_t^{\xi_a}(K,L) = A^{\xi_a}(K,L)\mathcal{R}_{t-1}^{\xi_a}(K,L) \oplus B^{d_a}(L)\mathcal{D}, \text{ for } t > 0 \tag{6.10}$$

where the subscript "$a$" to the variable (e.g., $\mathcal{R}_t^{\xi_a}(K,L)$, $A^{\xi_a}$, and $B^{d_a}$) indicates the presence of an attack. The minimum invariant set of the attacked closed-loop process with $(K, L)$ stable in the sense that $\rho(A^{\xi_a}(K,L)) < 1$, may be expressed (similar to Eq. 6.7) as:

$$\mathcal{R}_\infty^{\xi_a}(K,L) = \bigoplus_{i=0}^{\infty} A^{\xi_a}(K,L)^i B^{d_a}(L)\mathcal{D} \tag{6.11}$$

### 6.1.4 Class of Reachable Set-Based Detection Schemes

In the presence of an attack on the process, the augmented state values may deviate from their expected values for the attack-free process, meaning that an anomaly in the augmented state value, if detected, may indicate the presence of an attack. However, the augmented state cannot be measured, and standard anomaly detection schemes monitor a process based on values of a monitoring variable that are a function of the augmented state. Commonly used monitoring variables include the measured output $y = [C^x\ 0]\xi + [0\ I]d$, the estimated output $\hat{y} = [C^x\ \ -C^x]\xi$, or the residual, which may be defined as $r := y - \hat{y} = [0\ C^x]\xi + [0\ I]d$. The ability of a detection scheme to detect an attack may vary with the monitoring variable used. For example, detection of some attacks may be possible when the detection scheme uses the residual to monitor the process; however, these attacks may go undetected when the detection scheme uses the measured output. This realization motivated the formulation of a generalized monitoring variable in [48, 49] which may be expressed in terms of the measured and estimated outputs as:

$$\eta_t = H^y y_t + H^{\hat{y}} \hat{y}_t \tag{6.12}$$

where $H^y$ and $H^{\hat{y}}$ are design parameters that dictate the choice of the monitoring variable. The monitoring variable $\eta$ is a design parameter for a detection scheme of the form in Eq. 6.15, in the sense that each choice of the monitoring variable gives a different reachable set-based detection scheme. For example, if $H^y = [C^x\ 0]$ and $H^{\hat{y}} = [0\ I]$, then $\eta = y$, and if $H^y = [0\ C^x]$ and $H^{\hat{y}} = [0\ I]$, then $\eta = r$.

In view of the definitions of the augmented state and disturbance vectors, the monitoring variable may be expressed as a linear combination of the augmented state and the disturbance vector as follows:

$$\eta_t = \underbrace{[(H^y + H^{\hat{y}})C^x\ \ -H^{\hat{y}}C^x]}_{:=C^\eta} \xi_t + \underbrace{[0\ H^y]}_{:=D^\eta} d_t \tag{6.13}$$

For the closed-loop process with $(K, L)$ and an initial set of states $\mathcal{R}_0^\xi$, its monitoring variable values are contained within the $t$-step reachable sets for all $t \in \mathbb{Z}^+$. From Eq. 6.13, the $t$-step reachable set of the monitoring variable depends on the control parameters $(K, L)$, the initial set of states $\mathcal{R}_0^\xi$, and the set $\mathcal{D}$, and may be expressed as:

$$\mathcal{R}_t^\eta(\mathcal{R}_t^\xi(K, L)) = C^\eta \mathcal{R}_t^\xi(K, L) \oplus D^\eta \mathcal{D} \tag{6.14}$$

In this section, a class of detection schemes that track the evolution of the monitoring variable at each time step with respect to the reachable sets of the attack-free process at that time step is used [49]. The detection logic is given by:

$$\phi_t(\eta_t) = \begin{cases} 0, \ \eta_t \in \mathcal{R}_t^{\eta}(\mathcal{R}_t^{\xi}(K,L)) \\ 1, \ \eta_t \notin \mathcal{R}_t^{\eta}(\mathcal{R}_t^{\xi}(K,L)) \end{cases} \tag{6.15}$$

where $\phi_t(\eta_t)$ is the output of the detection scheme. An output of $\phi_t(\eta_t) = 1$ at time step $t \in \mathbb{Z}^+$ indicates that the detection scheme generates an alarm because an attack has been detected, while $\phi_t(\eta_t) = 0$ indicates that no alarm is raised because an attack has not been detected. For a concise presentation of the results, the class of detection schemes Eq. 6.15 will be referred to as the reachable set-based detection scheme. The reachable set-based detection scheme guarantees a zero false alarm rate during transient process operation because it accounts for the evolution of the monitoring variable of the attack-free process.

In [49], a method to classify attacks based on their detectability was presented, which was defined as the ability of the detection scheme in Eq. 6.15 to detect an attack. An overview of the detectability-based classification of attacks is presented here. If an attack on the closed-loop process with $(K, L)$ and the initial set of states $\mathcal{R}_0^{\xi}$ is detected by the detection scheme in finite time, the attack is defined as a detectable attack with respect to the detection scheme in Eq. 6.15. Attacks that cannot be detected by the detection scheme in finite time are defined as undetectable attacks with respect to the detection scheme in Eq. 6.15. Finally, with respect to the detection scheme in Eq. 6.15, a potentially detectable attack is defined as an attack that is neither detectable nor undetectable.

For the process under attack, the reachable sets for the monitoring variable are given by:

$$\eta_t^a = \underbrace{[(H^y\Lambda^y + H^{\hat{y}})C^x \ \ -H^{\hat{y}}C^x]}_{:=C^{\eta_a}}\xi_t + \underbrace{[0 \ H^y\Lambda^y]}_{:=D^{\eta_a}}d_t \tag{6.16a}$$

$$\mathcal{R}_t^{\eta_a}(\mathcal{R}_t^{\xi_a}(K,L)) = C^{\eta_a}\mathcal{R}_t^{\xi_a}(K,L) \oplus D^{\eta_a}\mathcal{D} \tag{6.16b}$$

It should be noted that, for the stable closed-loop system (both in the presence and in the absence of an attack), the monitoring variable is ultimately bounded within the minimum invariant set. The minimum invariant set of the monitoring variable is computed based on the minimum invariant set of the augmented state and the disturbance set, using the following relationship:

$$\mathcal{R}^{\eta}_{\infty}(K, L) = C^{\eta}\mathcal{R}^{\xi}_{\infty}(K, L) \oplus D^{\eta}\mathcal{D}$$
$$\mathcal{R}^{\eta_a}_{\infty}(K, L) = C^{\eta_a}\mathcal{R}^{\xi_a}_{\infty}(K, L) \oplus D^{\eta_a}\mathcal{D}$$

where $\mathcal{R}^{\eta}_{\infty}(K, L)$ and $\mathcal{R}^{\eta_a}_{\infty}(K, L)$ are the minimum invariant sets of the monitoring variable for the attack-free and attacked processes, respectively. In the remainder of the paper, to distinguish these minimum invariant sets from those associated with the augmented state, the minimum invariant sets for the monitoring variable are referred to as the terminal sets. From Eq. 6.14 and Eq. 6.16b, it can be seen that the detectability of an attack (with respect to the detection scheme in Eq. 6.15 is dependent on how the reachable sets of the monitoring variable for the process under attack evolve in relation to the reachable sets of the monitoring variable for the process in the absence of an attack. Given an attack and the set of initial states, the reachable sets for the attack-free and the attacked processes may be computed offline, and attacks may be classified on the basis of their detectability with respect to the reachable set-based detection schemes by checking for certain conditions [49]. If at any time step $t \in \mathbb{Z}^{+}$, the intersection between the reachable set for the attacked process and the reachable set for the attack-free processes is empty $(\mathcal{R}^{\eta}_{t}(\mathcal{R}^{\xi}_{t}(x^{s}_{f})) \cap \mathcal{R}^{\eta_a}_{t}(\mathcal{R}^{\xi_a}_{t}(x^{s}_{f})) = \emptyset)$, then the attack is detectable. An attack is undetectable if, for all $t \in \mathbb{Z}^{+}$, the reachable set for the attacked process is contained within the reachable set for the attack-free process, i.e., $\mathcal{R}^{\eta_a}_{t}(\mathcal{R}^{\xi_a}_{t}(x^{s}_{f})) \subseteq \mathcal{R}^{\eta}_{t}(\mathcal{R}^{\xi}_{t}(x^{s}_{f}))$. An attack is potentially detectable if for all $t \in \mathbb{Z}^{+}$, the reachable sets for the attacked process intersect with, but not necessarily contained within, the reachable sets for the attack-free process, i.e., $\mathcal{R}^{\eta_a}_{t}(\mathcal{R}^{\xi_a}_{t}(x^{s}_{f})) \cap \mathcal{R}^{\eta}_{t}(\mathcal{R}^{\xi}_{t}(x^{s}_{f})) \neq \emptyset$ for all $t \in \mathbb{Z}^{+}$ and $\mathcal{R}^{\eta_a}_{t}(\mathcal{R}^{\xi_a}_{t}(x^{s}_{f})) \nsubseteq \mathcal{R}^{\eta}_{t}(\mathcal{R}^{\xi}_{t}(x^{s}_{f}))$ for some $t \in \mathbb{Z}^{+}$.

## 6.2 Randomized Control Mode Switching for Detection of Cyberattacks

In this section, a reachable set-based detection scheme that employs randomized switching between different control modes to facilitate attack detection. To provide context for the development of the proposed scheme, is presented. First, a review of the previous work on control mode switching-enabled attack detection for processes under steady-state operation is presented. Following this, theoretical results that characterize the interdependence between the control parameter selection, the stability of the closed-loop process under attack, and the detectability of an attack with respect to the reachable set-based detection scheme in Eq. 6.15 are presented. Then, the algorithm for a randomized control mode switching-enabled attack detection for processes under transient operation to enable attack detection with zero false alarms is presented. Finally, a modification to the control mode switching algorithm that enables attack detection with zero false alarms when implemented on processes under steady-state operation is proposed.

### 6.2.1 Control Mode Switching for Cyberattack Detection

Multiplicative FDI attacks alter the stability properties of a closed-loop process by modifying the eigenvalues of the matrix $A^{\xi_a}(K, L)$ in Eq. 6.9. As a result, the detectability of such attacks may be influenced by the stability of the closed-loop process under an attack. In prior works [48, 50], a control mode switching-enabled attack detection method for the detection of multiplicative sensor–controller link FDI attacks was presented. Processes under steady-state operation were considered, for which the augmented state and the monitoring variable are bounded within the minimum invariant set, and the terminal set, respectively. The switching-enabled attack detection method enabled attack detection by exploiting the interdependence between the control parameter selection, the stability of the closed-loop process under an attack, and the detectability of an attack with respect to the terminal set-based detection scheme. Dynamics of the attack-free process may be excited due to a control mode switch and may cause a brief transient operation of the process during which the monitoring variable may not be bounded within the terminal set. The terminal set-based detection scheme may therefore generate a false alarm during transient operation, as it does not distinguish between anomalies in values of the monitoring variable during an attack and during transient operation. For processes with an invertible output matrix, a condition was presented that may be checked to schedule control mode switches at time steps when the switch does not cause a transient operation in the attack-free process [48]. However, satisfaction of the presented condition could not be guaranteed. To enable attack detection, a control mode switch could be implemented at a time when the condition is not satisfied, potentially leading to a transient operation, which could trigger false alarms in the detection scheme. Therefore, the proposed method did not eliminate false alarms.

In this section, the reachable set-based detection scheme in Eq. 6.15 is utilized to monitor a process during transient operation to eliminate false alarms even when the output matrix is not invertible. From Eq. 6.10, the reachable sets for the attacked process are influenced by the control parameters, and by extension, the control parameters influence attack detectability with respect to the reachable set-based detection scheme. If the attacked closed-loop process is unstable, its reachable sets may evolve differently from the reachable sets associated with the attack-free process, thereby influencing attack detectability. Proposition 11 below characterizes the interdependence between the stability of the closed-loop process under an attack and the detectability of an attack with respect to the detection scheme in Eq. 6.15.

**Proposition 11.** *Consider the attacked closed-loop process with* $(K, L)$ *and an initial set of states* $\mathcal{R}_0^{\xi}$. *Let an attack destabilize the process in the sense* $\|\xi_t\| \to \infty$ *as* $t \to \infty$. *If the matrix pair* $(A^{\xi_a}(K, L), C^{\eta_a})$ *is observable, then the attack is detectable with respect to the reachable set-based detection scheme in Eq. 6.15.*

Proposition 11 establishes a sufficient condition for an attack to be detectable by the reachable set-based detection scheme. To enable attack detection, it may be preferable to operate the closed-loop process under an "attack-sensitive" mode, in which the control parameters are chosen such that an attack destabilizes the process. However, as noted in [48, 50], prolonged operation of the process under the attack-sensitive control mode may

be undesirable because a tradeoff between attack detection and closed-loop performance may exist. To manage the tradeoff, extended operation of the process under a "nominal" control mode is considered, for which the control parameters $(K^i, L^i)$ are chosen to meet closed-loop performance considerations. For enabling attack detection, control mode switching is utilized to operate the process under the attack-sensitive control mode, for which the control parameters $(K^f, L^f)$ are chosen so that a range of attacks are detectable (per Proposition 11). Attack-sensitive control parameters are chosen so that an attack in the range considered causes the closed-loop process operated under the attack-sensitive control mode to be unstable in the sense that $\rho(A^{\xi_a}(K^f, L^f)) > 1$ and the matrix pair $(A^{\xi_a}(K^f, L^f), C^{\eta_a})$ is observable.
Control parameters influence the reachable sets of the attack-free process (from Eq. 6.6). Under the switching-enabled detection method, the control parameters may vary with the time step. Therefore, to monitor the switched system, the detection scheme in Eq. 6.15 is modified as follows:

$$\phi_t(\eta_t) = \begin{cases} 0, \ \eta_t \in \mathcal{R}^{\eta}_t(\mathcal{R}^{\xi}_t) \\ 1, \ \eta_t \notin \mathcal{R}^{\eta}_t(\mathcal{R}^{\xi}_t) \end{cases} \tag{6.18}$$

where $\mathcal{R}^{\eta}_t(\mathcal{R}^{\xi}_t)$ is the reachable set of the attack-free process that is computed per the following recurrence relation:

$$\mathcal{R}^{\xi}_t = A^{\xi}(K_{t-1}, L_{t-1})\mathcal{R}^{\xi}_{t-1} \oplus B^d(L_{t-1})\mathcal{D}, \ \text{for } t > 0 \tag{6.19}$$

where $(K_{t-1}, L_{t-1})$ are the gains at time $t-1$, $(K_t, L_t) = (K^N, L^N)$ if the control system is operated under the nominal mode, and $(K_t, L_t) = (K^A, L^A)$ if the control system is operated under the attack-sensitive mode. The dependence of the reachable sets for the switched system on the control parameters is dropped for conciseness, i.e., the reachable sets for the augmented state and the monitoring variable at any time step $t \in \mathbb{Z}^+$ is denoted simply as $\mathcal{R}^{\eta}_t(\mathcal{R}^{\xi}_t)$ and $\mathcal{R}^{\xi}_t$. The reachable sets of the attack-free process are initialized at the initial set of states ($\mathcal{R}^{\xi}_t = \mathcal{R}^{\xi}_0$ at $t = 0$) and assume that the first active mode is the nominal mode so that the first switch occurs from the nominal to the attack-sensitive mode. For attack detection, randomly choosing the control mode switching instance ($t_s$) may be preferred to scheduling the control mode switch at pre-determined time steps. This is because randomly choosing the switching instances may prevent an attacker from learning the switching schedule and designing a "smart" attack that evades detection. The reachable set-based detection scheme in Eq. 6.18 accounts for the evolution of the attack-free monitoring variable under each control mode, and guarantees a zero false alarm rate in the presence of any control mode switching implemented at any randomly chosen switching instance ($t_s \in \mathbb{Z}^+$). This is formally stated in the following proposition, which considers multiple control mode switches between the nominal and the attack-sensitive modes implemented on the attack-free process at randomly chosen time steps $t_{s_i} \in \mathbb{Z}^+$, where $i \in \{1, 2, 3, \ldots\}$ such that $t_{s_{i+1}} > t_{s_i}$. The odd values of $i$ represent the time steps at which a switch from the nominal to the attack-sensitive mode is implemented, while even values of $i$ represent the time steps when a switch back from the attack-sensitive to the nominal mode occurs.

**Proposition 12.** *Consider the attack-free closed-loop process under the nominal mode with an initial set of states $\mathcal{R}_0^{\xi}$, which is monitored by the reachable set-based detection scheme in Eq. 6.18. Let multiple control mode switches between the nominal and the attack-sensitive mode be implemented on the process. Let the switching instances be randomly chosen time steps $t_{s_i} \in \mathbb{Z}^+$, where $i \in \{1, 2, 3, \ldots\}$ such that $t_{s_{i+1}} > t_{s_i}$. The reachable set-based detection scheme generates no alarms for all $t \in \mathbb{Z}^+$.*

From Proposition 12, the reachable set-based detection scheme guarantees a zero false alarm rate under multiple successive switches between the nominal and attack-sensitive control modes implemented on the attack-free process. The absence of an attack on the process is a sufficient condition to be satisfied for zero false alarms, meaning that a lack of alarms may not be indicative of attack-free process behavior. However, it follows from Proposition 12 that if the reachable set-based detection scheme generates an alarm at any time step $t_d \in \mathbb{Z}^+$, then the alarm is only due to an attack on the process.

In general, it is not known if an attack is occurring on the process. To probe for an ongoing attack, multiple successive control mode switches may be implemented on the process. To facilitate the detection of a wide range of attacks, successive switches from the nominal control mode to multiple attack-sensitive control modes may be considered. However, the closed-loop process under the switching-enabled detection strategy is a switched system. If not executed carefully, multiple control mode switches may destabilize the attack-free switched closed-loop system. To avoid potential instability of the attack-free closed-loop system, a minimum dwell time approach can be used, whereby the process is forced to remain in each control mode for a minimum period of time before switching to another control mode. The minimum dwell time required for closed-loop stability can be characterized using Lyapunov techniques for switched systems [99]. Utilizing the minimum dwell time approach to ensure the stability of the switched closed-loop system means that each switching instance is lower bounded by the minimum dwell time of process operation under the previous control mode. As a result, the switching instances under the randomized switching-enabled detection strategy are random subject to a constraint on stability.

While the dwell time for the nominal control mode ($T_c^N$) may be chosen to maintain the stability of the attack-free switched closed-loop system, additional considerations may apply to the selection of a suitable dwell time for the process when operated under the attack-sensitive control mode ($T_c^A$). One consideration is management of the tradeoff between attack detection and attack-free closed-loop performance. A longer dwell time in the attack-sensitive mode may increase the chances of attack detection, but it could also degrade closed-loop performance in the absence of attacks. Meeting process safety constraints could also be another consideration in the selection of the dwell time for the attack-sensitive mode. In the presence of an attack, the augmented state under the attack-sensitive control mode may grow unbounded with time. However, the augmented state may remain bounded within some safe set for a finite time of operation under the attack-sensitive mode. It is possible that the considerations for the selection of a suitable dwell time for operation in the attack-sensitive mode may be conflicting in some cases. For example, to ensure that process safety constraints are met, the dwell time for the attack-sensitive mode may be shorter than the minimum dwell time required for the stability of the attack-free switched closed-loop system. In such cases, the dwell time for the attack-sensitive mode should be chosen based on the process safety constraints, while the closed-loop stability of the attack-free switched closed-loop system can be guaranteed by implementing only a finite number of control mode switches over the infinite time interval.

Further design considerations for the switching-enabled active detection method are the switching instances ($t_{s_i}$, $i \in \{1, 2, 3, \ldots\}$), which are randomly chosen time steps at which the control system switches from one control mode to the other. However, the stability considerations for the attack-free process may constrain control mode switching randomization. For switching from the nominal control mode to the attack-sensitive control mode, the switching instances are chosen as random integers such that the process is operated under the nominal mode for a time period that is at least equal to the dwell time for the nominal mode $T_c^N$. For switching from the attack-sensitive control mode to the nominal control mode, the switching instances are chosen such that the process is operated under the attack-sensitive control mode for the specified dwell time $T_c^A$ (chosen to meet closed-loop stability, closed-loop performance, and process safety constraints). The first switching instance is a random number that is greater than or equal to the dwell time of the process operated under the nominal control mode, i.e., $t_{s_1} \geq T_c^N$. All subsequent switching instances depend on the previous switching instance and account for the dwell time of the process under each mode. If no attack is detected, the even switching instances (for switching from the attack-sensitive control mode to the nominal control mode) may depend upon the prior switching instances and are chosen per the relation $t_{s_{2n}} = t_{s_{2n-1}} + T_c^A$ for $n \geq 1$. Similarly, the switching instances for switching from the nominal mode to the attack-sensitive mode are selected as random integers that account for the dwell time under the nominal control mode per the relation $t_{s_{2n+1}} \geq t_{s_{2n}} + T_c^N$ for $n \geq 1$.

Once an attack is detected, no further control mode switches are implemented in the process. From the results and discussion presented herein, monitoring the process using the reachable set-based detection scheme ensures that control mode switches can be randomized (randomly choose the switching instances $t_{s_i}$) to enable attack detection while guaranteeing a zero false alarm rate. Randomization of the control mode switches may help preserve the confidentiality of the detection scheme from an attacker engaged in espionage, and thereby prevent them from learning the switching schedule and designing a smart attack that could evade detection. Additionally, no assumptions on the structure of the output matrix ($C^x$) are made. Therefore, the switching-enabled detection method present in this section may enable attack detection with a guaranteed zero false alarm rate even when $C^x$ is non-invertible.

**Remark 6.2.1.** *To select the attack-sensitive control parameters, the range of attacks to be detected may be chosen as attacks on the process under the nominal mode which are either undetectable (attacks such that $\mathcal{R}_t^{\eta_a}(\mathcal{R}_t^{\xi_a}(K^N, L^N)) \subseteq \mathcal{R}_t^{\eta}(\mathcal{R}_t^{\xi}(K^N, L^N))$ for all $t \in \mathbb{Z}^+$) or those attacks which are potentially detectable (attacks such that $\mathcal{R}_t^{\eta_a}(\mathcal{R}_t^{\xi_a}(K^N, L^N)) \nsubseteq \mathcal{R}_t^{\eta}(\mathcal{R}_t^{\xi}(K^N, L^N))$ for some $t \in \mathbb{Z}^+$ and $\mathcal{R}_t^{\eta_a}(\mathcal{R}_t^{\xi_a}(K^N, L^N)) \cap \mathcal{R}_t^{\eta}(\mathcal{R}_t^{\xi}(K^N, L^N)) \neq \emptyset$ for all $t \in \mathbb{Z}^+$ ). More details on the selection of attack-sensitive parameters may be found in [50].*

**Remark 6.2.2.** *The even switching instances may be selected randomly to vary the dwell time of the process operated under the attack-sensitive mode between the minimum dwell time that allows for the stability of the attack-free switched system ($T_c^{A^{min}}$) and the specified dwell time chosen to meet the closed-loop performance, attack detection, stability, and process safety constraints ($T_c^A$). Under this modified method, the even switching instances may be chosen as random integers that satisfy $t_{s_{2n}} - t_{s_{2n-1}} \in [T_c^{A^{min}}, T_c^A]$ for all $n \geq 1$. However, implementing control mode switching with a time-varying dwell time for the attack-sensitive mode is subject to a rigorous characterization of the dwell times $T_c^{A^{min}}$ and $T_c^A$, which is outside the scope of the work presented in this chapter.*

### 6.2.2 Algorithms for the Randomized Control Mode Switching-Enabled Cyberattack Detection Method

Algorithm 2 outlines the steps for the implementation of the randomized control mode switching-enabled cyberattack detection method. The proposed method considers that the reachable sets of the monitoring variable for the attack-free process are computed

**Algorithm 2:** Algorithm for the randomized control mode switching-enabled attack detection with online reachable set computation

```
Inputs: $\mathcal{R}_0^\xi$, $(K^N, L^N)$, $(K^A, L^A)$, $T_c^N$, $T_c^A$
Initialization: $t_d = \infty$, $t_s = \infty$, $t = 0$, $(K_t, L_t) = (K^N, L^N)$
Outputs: $t_d$
do
    Receive the measured variable $y_t$ from the sensors over the sensor–controller
      communication link
    Calculate the monitoring variable $\eta_t$ and reachable set of the attack-free process
      $\mathcal{R}_t^\eta(\mathcal{R}_t^\xi)$
    Randomization logic
    if $t_s = \infty$ then
        Randomly generate switching flag : $f \in \{0, 1\}$
    else if $t > t_s + T_c^A + T_c^N$ then
        Set $t_s = \infty$
    Monitoring logic
    Compute the detection scheme output $\phi_t(\eta_t)$ per Eq. 6.18
    if $\phi_t(\eta_t) = 1$ then
        Declare the detection of an attack. Set the detection time step to be the
          current time step $t = t_d$. Terminate the detection algorithm.
    Switching logic
    else if $f = 1$ and $t_s = \infty$ then
        Set $f = 0$, $t_s = t$ and $(K_t, L_t) = (K^f, L^f)$
    else if $t = t_s + T_c^A$ then
        Set $(K_t, L_t) = (K^N, L^N)$
    Set $t \leftarrow t + 1$, $(K_{t+1}, L_{t+1}) = (K_t, L_t)$
while $t_d = \infty$;
```

online and allows for multiple switches between the nominal and the attack-sensitive control modes to probe the process for attacks. The inputs to the algorithm are the set of initial states $\mathcal{R}_0^\xi$, the nominal control parameters $(K^N, L^N)$, the attack-sensitive control parameters $(K^A, L^A)$, the minimum dwell time for the attack-free process operated in the nominal control mode $T_c^N$, and the dwell time of the closed-loop process operated in the attack-sensitive control mode $T_c^A$. The detection time $t_d$ is the output of the algorithm. The randomization of the control mode switching instances is dictated by the randomization flag $f$, which is a random variable that can take values of 0 or 1 and is assigned a new value at each time step (which is the controller sampling instance). If at a given time step, $f = 0$, then the current time step is not a switching instance, and if $f = 1$, then the current time step is a switching instance.

Algorithm 2 may be implemented with the reachable sets computed online using one of several methods that have been proposed in the literature (e.g., [90, 100, 101]). Online computation of the reachable sets may not scale well with the dimension of the state, and the computation may become intractable within the sampling instances. As a result, implementation of Algorithm 2 on a process under transient operation may not always be feasible.

Chemical processes are operated for extended periods at or near their steady states, where all possible values of the process states are bounded within the minimum invariant set. Considering this, a modification to the algorithm is proposed for the implementation of control mode switching on a closed-loop process under steady-state operation using reachable sets computed offline. Based on standard results from the literature (see Theorem 1 in [77]), it can be shown that after a switch between the nominal and attack-sensitive control modes, the reachable sets of the attack-free process converge to an invariant neighborhood of the minimum invariant set of the process under the new mode in finite time. This means that, for the process at steady state, the values of the monitoring variable evolve within the terminal set of the operating mode prior to a control mode switch. Following each switching event, for the process under transient operation, its monitoring variable values evolve within the attack-free reachable sets under the new mode. However, the transient operation lasts only for a finite number of time steps, until the reachable sets are contained entirely within (i.e., converge to) an invariant neighborhood of the minimum invariant set of the attack-free process operated under the new mode. After convergence, the process monitoring variable values are bounded within an invariant neighborhood of the terminal set of the attack-free process under the new mode. Based on these considerations, a hybrid approach may be used to monitor the switched closed-loop system. In this approach, the detection scheme switches with the control mode. The reachable set-based detection scheme is used only during the transient period after each switch. The detection scheme switches to a terminal set-based detection scheme after sufficient time has elapsed from the control mode switch such that the reachable sets of the augmented state for the attack-free process converge to an invariant neighborhood of the minimum invariant set of the process under the new mode. Using this hybrid monitoring approach, the computational load for computing the reachable sets during the implementation of Algorithm 2 may be reduced by terminating the online computation of the reachable sets after the process has attained steady-state operation (i.e., at the time step that the attack-free reachable sets converge to an invariant neighborhood of the minimum invariant set).

Algorithm 3 outlines the steps for implementing the randomized switching-enabled attack detection strategy for processes under steady-state operation that utilizes the hybrid monitoring approach. The inputs for the algorithm are the specified dwell time of the process operated under the nominal mode $T_c^N$, the dwell time of the process operated under the attack-sensitive mode $T_c^A$, the nominal $(K^N, L^N)$ and the attack-sensitive $(K^A, L^A)$ control parameters, the minimum invariant sets for the attack-free process operated under the nominal mode $\mathcal{R}_\infty^\xi(K^N, L^N)'$ and under the attack-sensitive mode $\mathcal{R}_\infty^\xi(K^A, L^A)'$, the number of time steps that the reachable sets take to converge to the minimum invariant set for the process operated under the attack-sensitive mode (starting from the minimum invariant set for the process operated under the nominal mode $t_r$), the number of time steps the reachable sets take to converge to the nominal minimum invariant set (starting from the attack-sensitive minimum invariant set $t_r^*$), and the reachable sets for a switch from the nominal control mode to the attack-sensitive control mode $\mathcal{R}_t^\eta(\mathcal{R}_t^\xi(K^A, L^A))$ for $t \in (0, t_r]$

**Algorithm 3:** Algorithm for the randomized control mode switching-enabled attack detection for processes with offline computation of reachable sets

```
Inputs: $\mathcal{R}^{\eta}_{\infty}(\mathcal{R}^{\xi}_{\infty}(K^N, L^N)')$, $\mathcal{R}^{\eta}_{\infty}(\mathcal{R}^{\xi}_{\infty}(K^A, L^A)')$, $T^A_c$, $T^N_c$,
        $\mathcal{R}^{\eta}_{t}(\mathcal{R}^{\xi}_{t}(K^A, L^A))$ for $t \in (0, t_r]$, $\mathcal{R}^{\eta}_{t}(\mathcal{R}^{\xi}_{t}(K^N, L^N))$ for $t \in (0, t^*_r]$
Initialization: $t = 0$, $t_d = \infty$, $t_s = \infty$, $\mathcal{R}^{\xi}_{0} = \mathcal{R}^{\xi}_{\infty}(K^N, L^N)'$, $(K_t, L_t) = (K^N, L^N)$
Outputs: $t_d$
while $t_d = \infty$ do
    Receive the measured variable $y_t$ from the sensors over the sensor-controller
      communication link
    Calculate the monitoring variable $\eta_t$
    Randomization logic
    if $t_s = \infty$ then
        Randomly generate switching flag : $f \in \{0, 1\}$
    else if $t > t_s + T^A_c + T^N_c$ then
        Set $t_s = \infty$
    Reachable sets for monitoring
    if $t_s = \infty$ or $t > t_s + t^*_r$ then
        $\mathcal{R}^{\eta}_{t}(\mathcal{R}^{\xi}_{t}) = \mathcal{R}^{\eta}_{\infty}(\mathcal{R}^{\xi}_{\infty}(K^N, L^N)')$
    else if $t \leq t_s + t_r$ then
        $\mathcal{R}^{\eta}_{t}(\mathcal{R}^{\xi}_{t}) = \mathcal{R}^{\eta}_{p}(\mathcal{R}^{\xi}_{t}(K^f, L^f))$
        $p = t + 1 - (t_s + t_r)$
    else if $t \in [t_s + T^A_c, t_s + T^A_c + t^*_r)$ then
        $\mathcal{R}^{\eta}_{t}(\mathcal{R}^{\xi}_{t}) = \mathcal{R}^{\eta}_{p}(\mathcal{R}^{\xi}_{t}(K^N, L^N))$
        $p = t + 1 - (t_s + T^A_c + t^*_r)$
    else if $t \in [t_s + t_r, t_s + T^A_c)$ then
        $\mathcal{R}^{\eta}_{t}(\mathcal{R}^{\xi}_{t}) = \mathcal{R}^{\eta}_{\infty}(\mathcal{R}^{\xi}_{\infty}(K^A, L^A)')$
```

**Algorithm 3:** Algorithm for the randomized control mode switching-enabled attack detection for processes with offline computation of reachable sets

```

    *Monitoring logic*
    Compute the detection scheme output $\phi_t(\eta_t)$ per Eq. 6.18
    **if** $\phi_t(\eta_t) = 1$ **then**
        Declare the detection of an attack.Set the detection time step to be the
          current time step $t = t_d$. Terminate the detection algorithm.
    *Switching logic*
    **else if** $f = 1$ *and* $t_s = \infty$ **then**
        Set $f = 0$, $t_s = t$ and $(K_t, L_t) = (K^A, L^A)$
    **else if** $t = t_s + T_c^A$ **then**
        Set $(K_t, L_t) = (K^N, L^N)$
    Set $t \leftarrow t + 1$, $(K_{t+1}, L_{t+1}) = (K_t, L_t)$
```

and for a switch from the attack-sensitive control mode to the nominal control mode $\mathcal{R}_t^{\eta}(\mathcal{R}_t^{\xi}(K^N, L^N))$ for $t \in (0, t_r^*]$. The algorithm implementation is terminated when an attack is detected.

**Remark 6.2.3.** *Per the steps in Algorithms 2 and 3, the odd switching instances, which dictate when a switch from the nominal control mode to attack-sensitive control mode occurs, may be chosen as an arbitrarily large positive integer. To avoid waiting for an inordinately long time period before a control mode switch, the time step for odd switching instances may be chosen as an integer over a finite time interval* $[t_s^{min}, t_s^{max}]$*, where* $t_s^{min} \geq T_c^N$ *is the lower bound and* $t_s^{max} \geq t_s^{min}$ *is the operator-specified upper bound of the interval.*

**Remark 6.2.4.** *When implementing Algorithm 2 and Algorithm 3, the randomization flag (*$f$*) may be drawn from a Bernoulli distribution with* $p \in [0, 1]$ *and* $q = 1 - p$*. Where,* $p$ *is the probability that a random variable drawn from the distribution takes a value of 1, and* $q$ *is the probability that a random variable from the distribution takes a value of*

*0. If $p = 0.5$, the process is switched as frequently as it is not switched. To increase the likelihood of confounding an attacker, $p$ may be chosen as a number that is not equal to 0.5.*

## 6.3 Application of Randomized Control Mode Switching Algorithms to Illustrative Processes

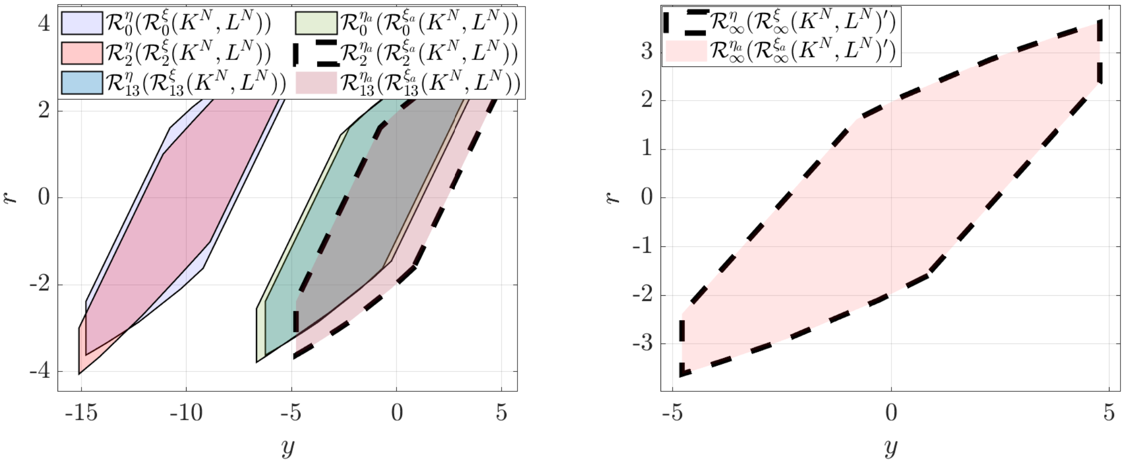
In this section, the application of the proposed switching-enabled attack detection strategy with randomized control mode switching is demonstrated using two illustrative examples. In the first example, the application of Algorithm 2 on a dynamic process is demonstrated, and the attack-free reachable sets are computed online for monitoring the process. The second example is a demonstration of the application of Algorithm 3 to a chemical process example under steady-state operation. In the second example, the attack-free reachable sets computed offline are used to monitor the process. The simulations of the chemical process example also demonstrate the application of the randomized control mode switching approach for the detection of a "smart" attack that is designed to evade detection under a scheduled control mode switch. The MPT 3.0 toolbox [82] is used to compute all polytopes [82] and the CORA toolbox [102] is used to compute all zonotopes.

### 6.3.1 Illustrative Scalar Process

A scalar process with a single state and a single manipulated input is considered:

$$\begin{aligned} x_{t+1} &= x_t + u_t + w_t \\ y_t &= \Lambda^y(x_t + v_t) \\ u_t &= \Lambda^u(-K\hat{x}_t) \end{aligned}$$

where $x_t \in \mathbb{R}$, $y_t \in \mathbb{R}$, and $u_t \in \mathbb{R}$ for all $t \in \mathbb{Z}^+$ are the process state, the measured output, and the manipulated input, respectively. The scalars, $v_t \in \mathcal{V} := \{v \in \mathbb{R} \mid |v| \leq 1\}$ and $w_t \in \mathcal{W} := \{w \in \mathbb{R} \mid |w| \leq 1\}$, model the bounded measurement noise and the process disturbance, respectively. The process may be under a multiplicative attack that modifies the operational data over all PCS communication links, which are represented by $\Lambda^y \neq 1$ and $\Lambda^u \neq 1$. To analyze the closed-loop process, an augmented state is defined that is a concatenation of the process state and the estimation error as follows: $\xi \coloneqq [x\ e]^T$. With this definition, the evolution of the augmented state of the process is described by Eq. 6.9 with $A^x = 1$, $B^u = 1$, $C^x = 1$. For process monitoring, a monitoring variable that is a concatenation of the measured output and the residual vectors $\eta \coloneqq [y\ r]^T$ is considered. The monitoring variable $\eta$ may be expressed in the form of in Eq. 6.12 with $H^y = [1\ 1]^T$ and $H^{\hat{y}} = [0\ -1]^T$.



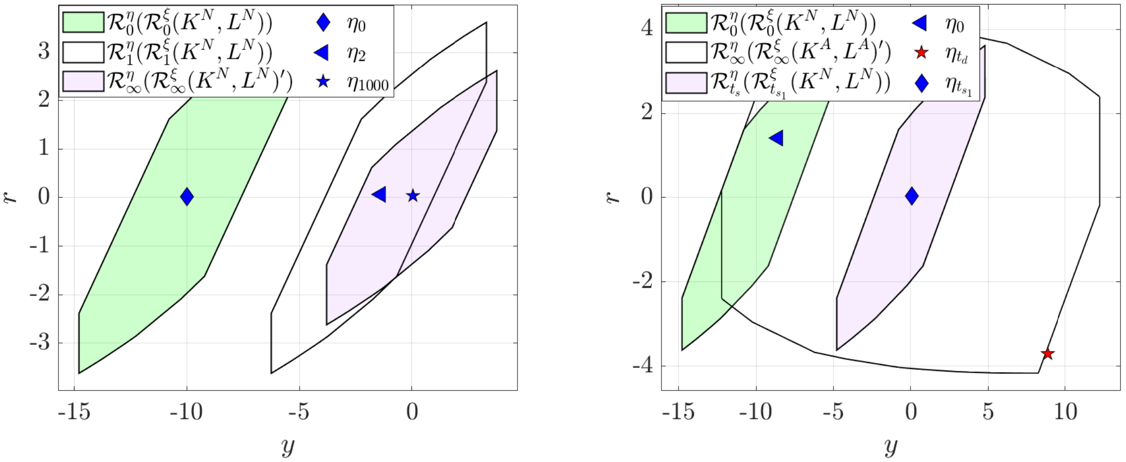
(a) (b)

Fig. 6.1: The reachable sets of the attacked and the attack-free process under the nominal mode (a) during transient operation when the attack is potentially detectable, and (b) at the steady-state when the attack is undetectable with $\mathcal{R}_{\infty}^{\eta_a}(\mathcal{R}_{\infty}^{\xi_a}(K^N, L^N)') \subset \mathcal{R}_{\infty}^{\eta}(\mathcal{R}_{\infty}^{\xi}(K^N, L^N)')$.

The nominal controller gain $K^N$ is chosen to minimize the quadratic cost $J = \mathbb{E}\big[\Sigma_{i=0}^{\infty}(x_i^T Q x_i + u_i^T R u_i)\big]$, with $Q = 5$, $R = 1$, and the nominal observer gain as the steady-state Kalman filter gain with covariance matrices $Q_L = 0.0011$ and $R_L = 0.0011$. Likewise, the attack-sensitive parameters are chosen such that the attacked closed-loop process operated under the attack-sensitive mode is unstable in the sense that $\rho(A^{\xi_a}(K^f, L^f)) > 1$ over the attack range $\Lambda^y = 0.86$ and $\Lambda^u \in [1.1, 2]$, and $\Lambda^u = 1.1$ and $\Lambda^y \in [0.1, 0.99]$. The matrix pair $(A^{\xi_a}(K^f, L^f), C^{\eta_a})$ is observable over the attack range considered for selecting the attack-sensitive parameters. The values of the nominal and attack-sensitive control parameters are $(K^N, L^N) = (0.8541, 0.618)$ and $(K^f, L^f) = (1.57, 1.28)$. For the attack-free process, invariant outer $\epsilon$-approximations of the minimum invariant set for the process under the nominal and the attack-sensitive modes are computed with an error bound of $\epsilon = 5 \times 10^{-5}$ using the method described in [79]. For brevity, in the remainder of this paper, the invariant outer $\epsilon$-approximation of the minimum invariant set is referred to as the minimum invariant set. For the attack-free process with the control system under the nominal mode, its reachable sets take 18 time steps to converge from the set of initial states to the minimum invariant set. Similarly, for the attack-free process operated exclusively under the attack-sensitive mode, the reachable sets of the augmented state from the set of initial states converge to the minimum invariant set in five time steps.

The process under a transient operation when its state evolves from a set of initial states is considered, that is, the polytope obtained by shifting all the vertices of the minimum invariant set of the attack-free process operated under the nominal mode by $\xi' = [-10\ 0]^T$. To quantify the performance of the controller for the attack-free process operated under the attack-sensitive mode and the nominal mode, two sets of simulations (each set consisting of 1000 simulations) of the attack-free process are performed. In the first set, the exclusive operation of the attack-free process under the nominal mode is considered. In the second set, the exclusive operation of the attack-free process under the attack-sensitive mode is considered. Within a simulation set, at each time step of each simulation, the values of the process disturbance and the measurement noise are varied. The process disturbances and measurement noise are modeled as random variables drawn from two separate normal distributions with $\mathcal{N}(0, 0.0333)$. However, across simulation sets, the same values of process disturbance and measurement noise are used. Within each simulation, the evolution of the process for 1000 time steps is considered, and the state is initialized at $\xi_0 = [-10\ 0]^T$. The quadratic cost ($J = \mathbb{E}\big[\Sigma_{i=0}^{1000}(x_i^T Q x_i + u_i^T R u_i)\big]$) across the simulation sets is compared. Over the simulations of the process operated under the nominal mode, the average quadratic cost was found to be 648.66 with a standard deviation of 37.59. Similarly, over simulations of the process operated under the attack-sensitive mode, the average quadratic cost was found to be 1189.68 with a standard deviation of 156.77. Comparing the performance of the controller between the two modes, it can be concluded that the controller performance is worse under the attack-sensitive mode. This result demonstrates that to manage the tradeoff between attack detection and the attack-free performance degradation resulting from control mode switching, intermittent switching from the nominal control mode to the attack-sensitive control mode may be preferred to operation under the attack-sensitive control mode exclusively.

Now, the switching-enabled detection method (Algorithm 2) is applied over simulations of the process during the transient operation for the detection of an attack with $\Lambda^y = 0.86$ and $\Lambda^u = 1.1$. First, the detectability of the attack is analyzed for when the process is operated under the nominal control mode by comparing the reachable sets associated with the attack-free process to those associated with the attacked process. Fig. 6.1a illustrates the reachable sets for the process operated under the nominal mode over a few time steps during transient operation. As illustrated, at time steps $t = 0, 2, 13$, the reachable sets for the attacked and the attack-free process always intersect; however, the attacked reachable sets are not contained within the attack-free reachable sets, meaning that the attack is potentially detectable. The transient operation of the attacked process lasts over 12 time steps over the time interval $t \in [0, 13)$. Over all time steps during transient operation of the process, the attack is found to be potentially detectable because the reachable sets satisfy $\mathcal{R}_t^{\eta_a}(\mathcal{R}_t^{\xi_a}(K^N, L^N)) \cap \mathcal{R}_t^{\eta}(\mathcal{R}_t^{\xi}(K^N, L^N)) \neq \emptyset$ and $\mathcal{R}_t^{\eta_a}(\mathcal{R}_t^{\xi_a}(K^N, L^N)) \not\subseteq \mathcal{R}_t^{\eta}(\mathcal{R}_t^{\xi}(K^N, L^N))$. At time step $t = 13$, the reachable sets of the attacked process converge to the terminal set of the attacked process. Therefore, to analyze attack detectability over the time steps $t \in (13, 18]$, the terminal set of the process under attack is compared with the reachable sets of the attack-free process, and the attack is found to be undetectable. Fig. 6.1b illustrates the terminal set of the attack-free process and the terminal set of the attacked process, showing that the attack on the process under steady-state operation is undetectable due to the fact that the terminal set of the attacked process is contained entirely within the terminal set of the attack-free process, i.e., $\mathcal{R}_t^{\eta_a}(\mathcal{R}_t^{\xi_a}(K^N, L^N)) \subset \mathcal{R}_t^{\eta}(\mathcal{R}_t^{\xi}(K^N, L^N))$. To verify attack detectability analysis, 1000 simulations of the attacked process operated exclusively under the nominal mode and monitored by the reachable set-based detection scheme in Eq. 6.15 are considered. Each simulation is initialized at $\xi_0 = [-10\ 0]^T$ and consider the evolution of the process states over 1000 time steps. Over all simulations, the values of the process disturbances and measurement noise are varied similar to the previous simulation sets. Over all simulations considering the process under an attack, the attack is not detected.



(a)

(b)

Fig. 6.2: Evolution of the monitoring variable values, with respect to the attack-free reachable sets, over a few time steps of a simulation of the scalar process with a control mode switch implemented at time step $t_{s_1} = 14$ for the case when (a) no attack takes place (demonstrating zero false alarms), and (b) an attack takes place at $t = 0$ and is detected at time step $t_d = 124$.

The switching-enabled active detection method is applied over simulations of the transient closed-loop process, and for monitoring, the reachable set-based detection scheme in Eq. 6.15 is used. Two sets of simulations (each consisting of 1000 simulations of the process) are designed similar to the simulations considered earlier when comparing the quadratic cost under the nominal and attack-sensitive modes. In the first simulation set, the attack-free process is considered, while in the second simulation set, the attacked process with the attack beginning at time step $t = 0$ is considered. In this section, detection of an attack on the process during transient operation is considered. Over each simulation, the switching-enabled attack detection method implements a single control mode switch at a randomly chosen time instance in the interval $[0, 17]$ when the attack-free process states under the nominal control mode are not within the minimum invariant set for the process operated under the nominal mode. Since the simulations consider a single switch between the two modes, no minimum dwell time for each mode is specified. Similarly, over simulations that consider the process operated under the attack-sensitive control mode, a dwell time of $T_c^A = 150$ is used. Across simulations of the attack-free and the attacked processes, the same switching instance is considered. To implement the detection scheme, the attack-free reachable sets are computed online at each time step by using Eq. 6.14 with $(K, L)$ selected based on the process operation mode (under the nominal mode $(K, L) = (K^N, L^N)$, while under the attack-sensitive mode $(K, L) = (K^A, L^A)$). However, to reduce the computational load, the online computation of the reachable sets is terminated at the time step when the attack-free augmented state is expected to be contained within the minimum invariant set for the process operated under the mode of operation considered. After termination of online computation of the reachable sets, the terminal set of the monitoring variable for the attack-free process under the mode of operation considered is used to monitor the process.

Over the 1000 simulations that consider the attack-free process, a switch from the nominal to the attack-sensitive control mode is implemented at time steps chosen randomly over the time interval $[0, 17]$. No attack detection occurs over all simulations after the first switch from the nominal control mode to the attack-sensitive control mode, and the control mode switches back to the attack-sensitive mode after 150 time steps from the first switch. No false alarms were observed over all simulations. Under the second simulation set, detection of the attack occurs over all simulations within a minimum of 5 time steps and a maximum of 135 time steps from the control mode switch. Over all simulations, the control system switches back to the nominal control mode, after which the attack is not detected. Fig. 6.2a and Fig. 6.2b show the values of the monitoring variable over a few time steps of one simulation considering the attack-free process, and one simulation considering the attacked process, respectively. Over both simulations, the control parameters switch from the nominal to attack-sensitive values at time step $t_{s_1} = 14$.

Fig. 6.2a shows the values of the monitoring variable values observed when the switch is implemented over a simulation considering the attack-free process. At time step $t = 0$, no false alarm is generated as the value of the monitoring variable at that time step represented by the blue diamond marker is contained within the attack-free reachable set at that time step, which is the set shown in green. Similarly, no false alarm is observed at time step $t = 1$ because the monitoring variable value represented as the blue diamond is contained within the attack-free reachable set at that time shown as the set in white. While omitted for clarity, no false alarms are observed until the control mode switch at $t_{s_1} = 14$ because the monitoring variable values at each time step evolve within the corresponding attack-free reachable sets. After the switch is implemented, the process is operated under the attack-sensitive mode for 150 time steps, during which no false alarms are observed. At time step $t_{s_1} + T_c^A = 164$, an attack is not detected, causing the control system to switch back to the nominal mode. No false alarms are observed even after this switch until the end simulation, at time step $t = 1000$, when the monitoring variable value represented by the blue star marker is contained within the attack-free terminal set shown as the set in purple.

Fig. 6.2b shows the values of the monitoring variable over a few time steps of a simulation considering the process under the attack. At time step $t = 0$, the attack on the process under the nominal mode is not detected because the monitoring variable value represented by the blue triangle marker is contained within the attack-free reachable set at that time step, shown as the green set. While omitted for clarity, the attack is not detected during the process operation under the nominal mode because the monitoring variable values over the time interval $t \in [0, 14)$ are contained within the corresponding attack-free reachable sets. After a switch from the nominal control mode to the attack-sensitive control mode, the attack is detected the time step $t_d = 124$ because the monitoring variable value represented by the red star marker leaves the attack-free reachable set (which is the terminal set of the process operated under the attack-sensitive control mode) at that time step shown as the white set. In this case, attack detection occurs after the reachable sets of the attack-free process converge to its terminal set under the attack-sensitive mode. After the attack is detected, the control systems switches back to the nominal control mode and no further alarms are observed because the monitoring variable values at each time step are

Process parameters of the CSTR

| | |
|---|---|
| Volumetric flow rate ($F$) | $5.0\,\mathrm{m^3\,h^{-1}}$ |
| Reactor volume ($V$) | $1.0\,\mathrm{m^3}$ |
| Feed concentration of $A$ ($C_{A0}$) | $4.0\,\mathrm{kmol\,m^{-3}}$ |
| Activation energy ($E$) | $5.0 \times 10^4\,\mathrm{kJ\,kmol^{-1}}$ |
| Pre-exponential factor ($k_0$) | $8.46 \times 10^6\,\mathrm{m^3\,h^{-1}\,kmol^{-1}}$ |
| Gas constant ($R$) | $8.314\,\mathrm{kJ\,kmol^{-1}\,K}$ |
| Feed temperature ($T_0$) | $300\,\mathrm{K}$ |
| Density of reactor liquid hold-up ($\rho$) | $1000\,\mathrm{kg\,m^{-3}}$ |
| Heat of reaction ($\Delta H$) | $-1.15 \times 10^4\,\mathrm{kJ\,kmol^{-1}}$ |
| Heat capacity ($C_p$) | $0.231\,\mathrm{kJ\,kg\,K^{-1}}$ |
| Steady-state heat rate added/removed from the reactor ($Q_s$) | $0\,\mathrm{kJ\,h^{-1}}$ |
| Steady-state reactant concentration ($C_{As}$) | $1.22\,\mathrm{kmol\,m^3}$ |
| Steady-state temperature ($T_s$) | $438.2\,\mathrm{K}$ |

contained within the corresponding attack-free reachable sets until the end of simulation when the attack-free reachable sets have converged to the terminal set of the attack-free process under the nominal control mode. The monitoring variable value at time step $t = 1000$ (end of the simulation) is represented by the blue diamond marker, and as shown, no alarm is observed at this time step because the monitoring variable is contained within the attack-free terminal set shown as the set in purple. These results demonstrate that the switching-enabled attack detection method utilizing the reachable set-based detection scheme enables attack detection on a dynamic process while guaranteeing a zero false alarm rate due to a control mode switch implemented at a randomly chosen time step.

### 6.3.2 A Continuous Stirred-Tank Reactor

An example process that consists of a continuous stirred tank reactor (CSTR) with a second-order exothermic reaction of the form $A \rightarrow B$ is considered. The process dynamics are modeled by the following system of ordinary differential equations:

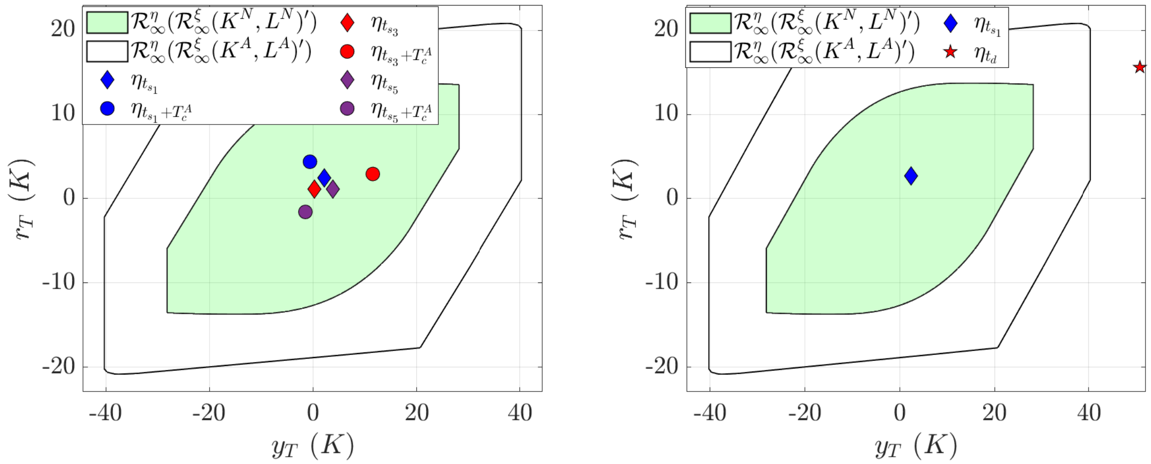
$$\begin{aligned}\frac{dC_A}{dt} &= \frac{F}{V}(C_{A0} + \Delta C_{A0} - C_A) - k_0 e^{\frac{-E}{RT}} C_A^2 \\ \frac{dT}{dt} &= \frac{F}{V}(T_0 + \Delta T_0 - T) - \frac{\Delta H k_0}{\rho C_p} e^{\frac{-E}{RT}} C_A^2 + \frac{Q}{\rho C_p V}\end{aligned} \tag{6.21}$$

where $C_{A0}$ and $T_0$ are the inlet reactant concentration and feed temperature, respectively, and $C_A$ and $T$ are the reactant concentration and reactor temperature, respectively. The rate of heat transfer to or from the reactor $Q$ is chosen as the manipulated input. The process is subject to bounded disturbances, modeled as variations in the inlet reactant concentration $\Delta C_{A0}$ and variations in the feed temperature $\Delta T_0$. The bounded process disturbances are within the limits $|\Delta C_{A0}| \leq 0.015\,\mathrm{kmol\,m^{-3}}$ and $|\Delta T_0| \leq 4.5\,\mathrm{K}$. The measured variable available to the controller is the reactor temperature $T$. The bounded noise in the measurements from the sensor is within limits such that $|v| \leq 4.5\,\mathrm{K}$. Table 2.1 provides a list of the definitions and values of the process parameters, the table is reproduced in this chapter to make it self-contained. Because the measurement of all possible states of the reactor are not available to the controller, the illustrative example presented in this section considers a case where the output matrix $C$ is non-square and non-invertible.

To obtain a model similar to Eq. 6.1, the continuous-time nonlinear process model in Eq. 6.21 is discretized using a sampling interval of $\Delta = 1 \times 10^{-2}$ h. The system matrices for the linearized CSTR process are:

$$A^x = \begin{bmatrix} 0.7364 & -0.0041 \\ 10.6953 & 1.156 \end{bmatrix},\ B^u = \begin{bmatrix} -0.0009 \times 10^{-4} \\ 0.4674 \end{bmatrix},\ B^w = \begin{bmatrix} 0.0433 & -0.001 \\ 0.2724 & 0.054 \end{bmatrix}$$

The nominal observer gain is chosen as the steady-state Kalman filter gain with covariance matrices, $Q_K = \begin{bmatrix} 1.4062 \times 10^{-5} & 0 \\ 0 & 1.2656 \end{bmatrix}$, $R_K = 1.2656$ and the nominal controller gain is chosen to minimize the quadratic cost $J = \mathbb{E}\left[\Sigma_{i=0}^{\infty}(x_i^T Q x_i + u_i^T R u_i)\right]$ with $Q = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}$ and $R = 10$. The attack-sensitive control parameters are chosen so that the attacked closed-loop process under the attack-sensitive mode is unstable under a range of attacks by checking if $\rho(A^{\xi_a}(K^f, L^f)) > 1$ over the attack range $\Lambda^y = 1.1$, $\Lambda^u \in [1.3, 2]$, and $\Lambda^u = 1.3$, $\Lambda^y \in [1.1, 2]$. Over the attack range considered, the matrix pair $(A^{\xi_a}(K^f, L^f), C^{\eta_a})$ is observable, indicating that all attacks in the range considered are potentially detectable.



(a)

(b)

Fig. 6.3: Evolution of the monitoring variable values observed over a few time steps of (a) the attack-free process with consecutive switches from the nominal to attack-sensitive mode implemented at $t_{s_1} = 221$, $t_{s_3} = 767$, and $t_{s_5} = 1313$ showing no false alarms due to switching and, (b) the attacked process with control mode switch from the nominal to the attack-sensitive mode implemented at time step $t_{s_1} = 221$ leading to attack detection at time step $t_d = 235$.

In this section, the switching-enabled detection method is applied on the CSTR process under steady-state operation, when all values of the process states are bounded within the minimum invariant set of the process. The disturbance set is modeled as a zonotope with the origin as the center. To compute invariant outer $\epsilon$-approximations of the minimum invariant set (henceforth referred to as the minimum invariant set for brevity) of the attack-free process under the nominal and the attack-sensitive modes, the method in [103] is used with an error bound of $\epsilon = 5 \times 10^{-5}$. For a switch performed on the attack-free process from the nominal control mode to the attack-sensitive control mode, the reachable sets are computed with the set of initial states $\mathcal{R}^{\xi}_0 = \mathcal{R}^{\xi}_{\infty}(K^N, L^N)'$ for $t_r = 11$ time steps when the sets are contained entirely within the minimum invariant set of the attack-free process under the attack-sensitive mode. Similarly, for a switch from the attack-sensitive control mode to the nominal control mode, the reachable sets are computed for the attack-free process with the set of initial states $\mathcal{R}_0^{\xi} = \mathcal{R}_{\infty}^{\xi}(K^A, L^A)'$ until they converge to the minimum invariant set under the nominal control mode after $t_r^* = 150$ time steps. The monitoring variable vector $\eta = [y\ r]^T$ is used, its terminal set under each control mode and the reachable sets after each control mode switch are computed until they are contained entirely within the terminal set under the new mode.

Initially, two simulation sets are performed to compare the quadratic cost $(J = \mathbb{E}\big[\Sigma_{i=0}^{1500}(x_i^T Q x_i + u_i^T R u_i)\big])$ for operating the process exclusively under the nominal mode with the cost of operation of the system exclusively under the attack-sensitive mode. The first set of simulations consider the attack-free process operated exclusively under the nominal mode, while the second set of simulations considers the exclusive operation of the attack-free process under the attack-sensitive mode. Under each set, 1000 simulations considering the process operating under attack-free conditions are conducted. Each simulation considers the evolution of the process over 1500 time steps spanning 15 h in real time. The process disturbances on the feed concentration are modeled as random variables drawn from a distribution with $\mathcal{N}(0, 0.005)$, and the process disturbance and measurement noise on the feed temperature and the temperature of the reactor (the measured output) are modeled as random variables picked from two distinct normal distributions with $\mathcal{N}(0, 1.5)$. Over each time step of each simulation, the values of the random variables representing the process distribution and the measurement noise are varied. However, the same values of random numbers are used across simulation sets. The disturbances are clipped at the absolute value of their bound to ensure that there are no false alarms (e.g., if the absolute value random number representing $\Delta C_{A0}$ exceeds 0.015, it is set to 0.015). Each simulation is initialized at the origin, which is contained within the minimum invariant sets of the process under the attack-sensitive and the nominal modes. Over all simulations, considering the process operated under the nominal mode, the quadratic cost has a mean of $1.016 \times 10^4$ and a standard deviation of $2.498 \times 10^3$. However, for the process operated exclusively under the attack-sensitive mode, the quadratic cost has a mean of $1.1985 \times 10^{15}$ and a standard deviation of $6.2674 \times 10^{13}$. Therefore, the performance of the controller under the attack-sensitive mode is higher than under the nominal mode, indicating that switching may be preferable to extended process operation under the attack-sensitive control mode.

Next, two simulation sets of the process monitored by the reachable set-based detection scheme in Eq. 6.15 are performed. In the first set, the Algorithm 3 is implemented over simulations considering the process without an attack to show that there are no alarms generated from consecutive control mode switches implemented at randomly chosen time steps. To this end, the detection of an attack with $\Lambda^y = 1.1$ and $\Lambda^u = 1.3$ is considered. The attack beginning at time step $t = 0$ on the process operated only under the nominal mode is not detected over 1000 simulations, even though it is potentially detectable. Therefore, in the second simulation set, the Algorithm 3 is implemented over 1000 simulations of the attacked process (attack begins at $t = 0$). Over simulations considering the process operated under the attack-sensitive mode, the dwell time under attack-sensitive mode is restricted to $T_c^A = 150$. Over each simulation, the switching instances are chosen randomly such that a maximum of three switches from the nominal to the attack-sensitive control mode and back from the attack-sensitive to the nominal control mode are possible over each simulation. Specifically, the first switching instance from the nominal to the attack-sensitive control mode ($t_{s_1}$) is selected as a random integer in the interval $[0, 1200]$. The second switching instance from the nominal control mode ($t_{s_3}$) is based on the first switching instance by selecting a random integer over the interval $[t_{s_1} + T_c^A + 150, 1200]$. If no attack is detected, then a third instance of a switch from the nominal control mode to the attack-sensitive control mode is allowed, with the switching instance $t_{s_5}$ chosen as a random integer over the interval $[t_{s_3} + T_c^A + 150, 1200]$. No minimum dwell time is specified for operation under the nominal mode, since the simulations consider a finite number of switches between the different control modes.

Over all simulations considering the attack-free process, no alarms are observed, and a minimum of two and a maximum of three control mode switches are implemented. Over all simulations, the process is under the nominal mode at the end of the simulation. Fig. 6.3a illustrates the values of the monitoring variable observed over a few time steps of a simulation of the attack-free process over which there are three consecutive switches from the nominal to the attack-sensitive control mode implemented at time steps $t_{s_1} = 221$, $t_{s_3} = 767$, and $t_{s_5} = 1313$. Over all simulations of the attack-free process, no false alarms are observed. While omitted for clarity in Fig. 6.3a, the monitoring variable values before the time step $t_{s_1}$ evolve within the terminal set of the attack-free process under the nominal control mode shown as the green set and no false alarms are observed until the first switching instance $t_{s_1} = 221$ when the control system switches to the attack-sensitive control mode. No false alarms are observed during process operation under the attack-sensitive control mode. As a result, after the dwell time under the attack-sensitive mode elapses, the control system switches back to the nominal mode at time step $t_{s_1} + T_c^A$. As shown, no false alarm is observed at $t_{s_1} + T_c^A$ because the monitoring variable value (indicated by the blue circle marker) is contained within the terminal set of the attack-free process under the attack-sensitive control mode (indicated by the white set). While omitted here for brevity, no false alarms are observed during process operation under the nominal control mode, and a second switch to the attack-sensitive control mode occurs at time step $t_{s_2} = 767$, when the monitoring variable value (indicated by the red diamond marker) is contained within the attack-free terminal set. Even after the second control mode switch, no false alarms are observed, and the control system switches back to the nominal control mode after the dwell time under the attack-sensitive mode elapses at time step $t_{s_2} + T_c^A$, when the monitoring variable value (indicated by the red circle marker) is contained within the terminal set of the process under attack-sensitive mode. No false alarms are observed even after a third switch from the nominal to the attack-sensitive control mode occurs at $t_{s_3}$ (monitoring variable value shown by purple diamond marker) followed by a switch back from the attack-sensitive to the nominal control mode at time step $t_{s_3} + T_c^A$ (monitoring variable value shown by purple circle marker). The results demonstrate that the proposed control mode switching strategy guarantees a zero false alarm rate when implemented on the attack-free process.

Fig. 6.3b illustrates the values of the monitoring variable observed over some time steps of one simulation of the process under the attack, over which the first switching instance is $t_{s_1} = 221$. Until this control mode switch, the attack is not detected because the monitoring variable values evolve within the terminal set of the attack-free process under the nominal control mode. At the switching instance $t_{s_1}$, the monitoring variable value shown by the blue marker is contained within the terminal set of attack-free process under the nominal mode (shown as the green set), meaning that no attack is detected. After the control mode switch, no alarms are observed until the attack is detected at time step $t_d = 235$ when the reachable sets of the attack-free process are contained entirely within the terminal set of the attack-free process operated under the attack-sensitive control mode. As shown in Fig. 6.3b, at the detection time step $t_d$, the monitoring variable value of the process shown by the red diamond marker is not contained within the terminal set of the attack-free process operated under the attack-sensitive control mode. Over all simulations of the process under attack, detection of the attack occurs after the first control mode switch within a minimum of 4 time steps and a maximum of 69 time steps from the switching instance. The results demonstrate that the reachable set-based detection scheme in Eq. 6.15 guarantees attack detection with a zero false alarm rate, for a randomly chosen switching instance, even when the output matrix $C^x$ is non-invertible.

#### 6.3.2.1 Comparison of Randomized and Scheduled Control Mode Switching

This section demonstrates the application of the randomized control mode switching to enable the detection of a "smart" attack that is designed to evade detection under a scheduled control mode switching-enabled attack detection method. A simulation of the CSTR process is considered, over which control mode switching is implemented per a fixed schedule as shown in Fig. 6.4. As illustrated, the switching schedule allows for two control mode switches between the nominal and the attack-sensitive control modes. For switching from the nominal to the attack-sensitive control mode, the first switching instance is $t_{s_1} = 300$ and the second switching instance is $t_{s_3} = 900$. If no attack is detected on the process operated under the attack-sensitive control mode until the dwell time of $T_c^A = 150$ time steps elapses, the controller switches back to the nominal mode. It is assumed that an attacker who is aware of the switching schedule designs a smart attack that switches the attack matrices in sync with the control mode switching. Specifically, the attacker uses an attack with $\Lambda^y = 1.1$ and $\Lambda^u = 1.3$ when the process is expected to operate under the nominal control mode. While potentially detectable, the attack on the process operated under the nominal control mode is not detected. However, when the process is expected to operate under the attack-sensitive control mode, the attacker switches to using an attack with $\Lambda^y = 0.9$ and $\Lambda^u = 1.1$ under which the attacked process is stable. Similar to the attack on the process operated under the nominal mode, the attack on the process operated under the attack-sensitive mode is potentially detectable; however, the attack is such that it is not detected. The attack schedule is as shown in Fig. 6.5a and Fig. 6.5b.

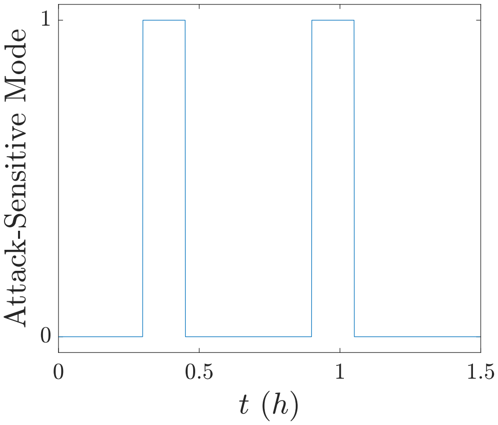


Fig. 6.4: Scheduled control mode switching for the CSTR process. A value of 0 on the Y-axis indicates that the process is operated under the nominal mode, and a value of 1 on the Y-axis indicates that the process is operated under the attack-sensitive mode.

One simulation is performed that considers the closed-loop process under the smart attack, with the control mode switch implemented per the schedule shown in Fig. 6.4. The process is initialized under the nominal control mode, with its initial state chosen as the origin. The reachable set-based detection scheme is used to monitor the process, similar to the previous section. The values of the process disturbances and measurement noise are varied at each time step over this simulation. The values of the monitoring variables at the switching instances over this simulation are illustrated in Fig. 6.6a. The attack is not detected over this simulation, demonstrating that an attacker with sufficient knowledge of the detection scheme may be able to design an attack that is capable of evading detection under a scheduled control mode switch.

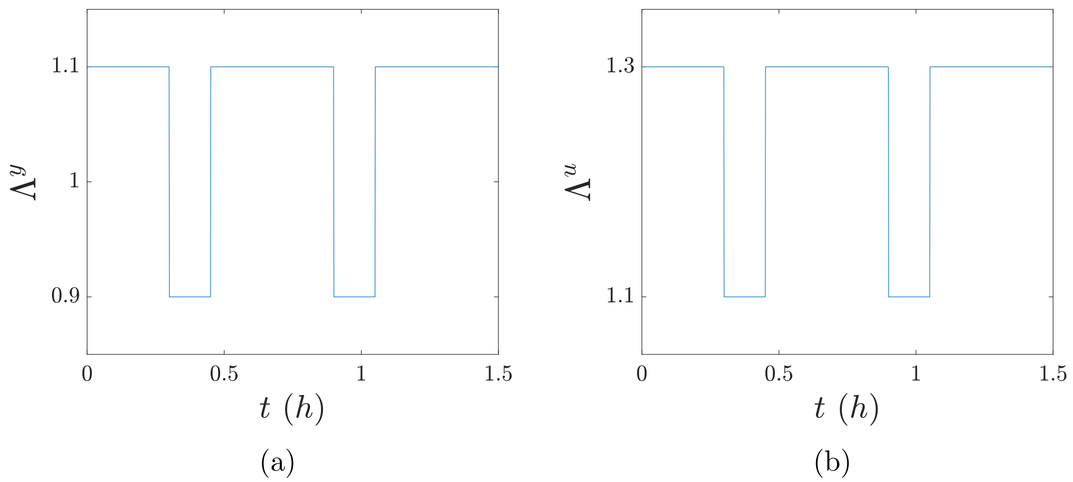
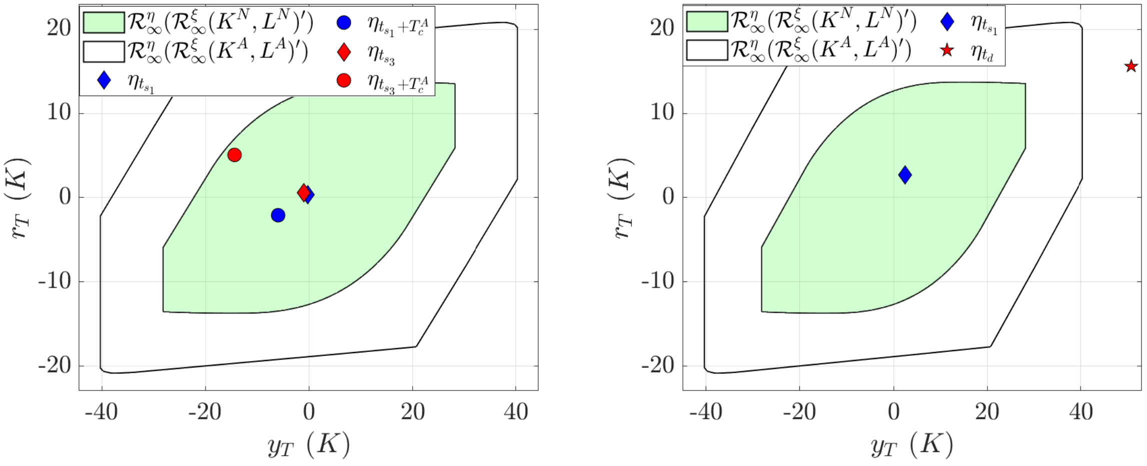


Fig. 6.5: The attack schedule for the smart attack, which is designed to evade detection under the scheduled control mode switching-based detection scheme. The figures illustrate the attack schedule for: (a) the sensor-controller attack, and (b) the controller-actuator attack.

Next, 1000 closed-loop simulations are performed considering the process subject to the smart attack but with randomized (rather than scheduled) control mode switching. Over each simulation, the same values of the process disturbances and measurement noise as in the first simulation with the scheduled control mode switch are used. However, the control mode switches are applied at randomly chosen time steps. Over each simulation, a maximum of three control mode switches is allowed. The first control mode switching instance from the nominal to the attack-sensitive control mode ($t_{s_1}$) is selected as a random integer generated over the interval $[99, 249]$. The second switching instance from the nominal to the attack-sensitive control mode ($t_{s_3}$) depends on the first switching instance and is selected as a random integer generated over the interval $[t_{s_1} + T_c^A + 151, t_{s_1} + T_c^A + 451]$. Finally, if an attack is not detected over the two previous control mode switches, a third switch from the nominal control mode to the attack-sensitive control mode is implemented at time step ($t_{s_5}$) selected as a random integer generated over the interval $[t_{s_3} + T_c^A + 151, t_{s_3} + T_c^A + 451]$. The third switch from the nominal to the attack-sensitive control mode is implemented only if $t_{s_3} + T_c^A + 151 < 1200$ so that at the end of each simulation, the process is under the nominal control mode.



(a)

(b)

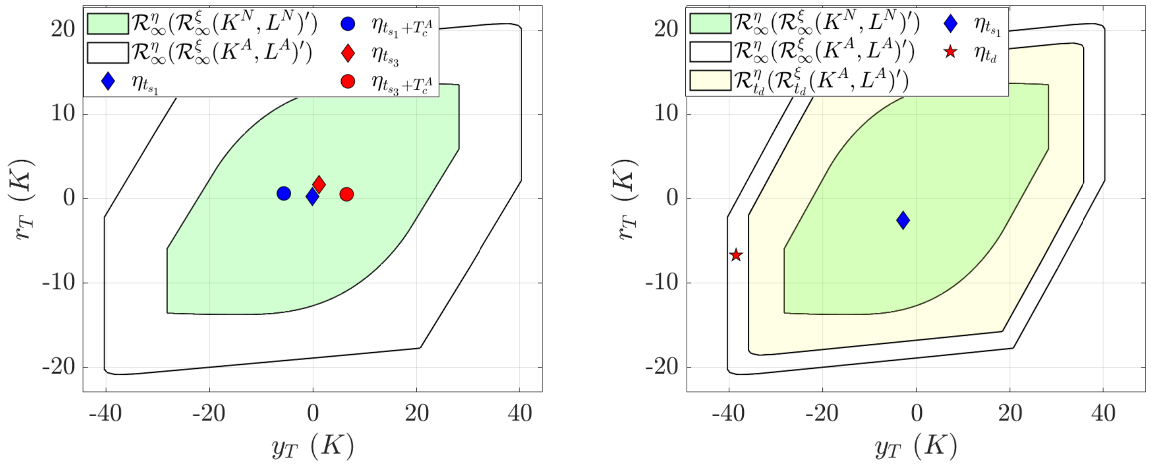
Fig. 6.6: Values of the monitoring variable for the CSTR process under attack showing (a) no attack detection with scheduled control mode switching (b) attack detection at $t_d = 235$ under the randomized switching scheme after a control switch implemented at $t_{s_1} = 221$.

Over all simulations considering the process under the attack and the randomized control mode switching, the detection of the smart attack occurs after the first control mode switch within a minimum of 4 time steps and a maximum of 40 time steps from the switching instance. Fig. 6.6b illustrates the attack detection over a simulation considering the process under the smart attack with the randomized control mode switching-enabled attack detection method. Over this simulation, the control mode switch is implemented at time step $t_{s_1} = 221$ leading to attack detection at $t_d = 235$. Because an attack is detected after the first control mode switch, no further switches are implemented by the detection scheme. The results demonstrate that from the perspective of enabling attack detection, a randomized control mode switching method may be preferred to a scheduled control mode switching method because an attacker may not be able to design a smart attack that is capable of evading detection.

**Remark 6.3.1.** *Fig. 6.4 is an illustrative example for the operation of the CSTR process under the scheduled switching-enabled detection strategy. An operator may choose to implement a switching-enabled detection strategy on the process that uses other periodic or* *non-periodic patterns for switching between the nominal and the attack-sensitive modes. Irrespective of the pattern of operation, an attacker with knowledge of the switching schedule may be able to design a detection-evading smart attack. Comparison of attack detection between scheduled and randomized switching-enabled strategies for smart attacks designed to evade detection under scheduled switching with patterns other than the one shown in Fig. 6.4 may require a case-by-case analysis. Nevertheless, it is anticipated that operating the CSTR process under the randomized switching-enabled strategy will help preserve the confidentiality of the detection scheme and thus enable attack detection over more simulations than under the scheduled switching-enabled detection strategy.*

#### 6.3.2.2 Application to the Nonlinear CSTR Process

The application of the proposed switching-enabled attack detection method to detect a smart attack on the nonlinear sampled-data model of the CSTR process is demonstrated. In all simulations considered in this section, the CSTR is modeled using its continuous-time nonlinear process model in Eq. 6.21, and the linear control law and the Luenberger observer are applied with a zero-order hold that considers a sampling interval of $\Delta = 1 \times 10^{-2}$ h. To solve the differential equations modeling the CSTR, the explicit Euler's method is used with an integration time step of $1 \times 10^{-4}$ h. The process disturbances and measurement noise values are modeled as random numbers drawn from a Gaussian distribution. Specifically, the disturbance in the feed concentration ($\Delta C_{A0}$) is modeled as a random variable drawn from a distribution with $\mathcal{N}(0, 0.0037)$ and the disturbance in the sensor measuring the temperature of the feed to the reactor and the measurement noise in the sensor measuring the temperature of the reactor is modeled as random variables drawn from two distinct distributions with $\mathcal{N}(0, 1.125)$. The disturbances are drawn from normal distributions with a smaller standard deviation than those used over the simulations in the Section 6.3.2.1. This is performed to ensure the validity of the reachable and terminal sets computed using the linear process model and enable process monitoring using the detection scheme in Eq. 6.15.



(a) (b)

Fig. 6.7: The values of the monitoring variable for the nonlinear CSTR under a smart attack showing: (a) no attack detection with a scheduled control mode switch, and (b) attack detection at time step $t_d = 231$ (2.31 h in real time) under the randomized switching scheme after a control mode switch at time step $t_{s_1} = 221$ (2.21 h in real time).

Similar to the previous section, first the closed-loop process under a smart attack (Fig. 6.5a and Fig. 6.5b) is simulated with a scheduled control switching-enabled attack detection method as in Fig. 6.4. Over this simulation, the closed-loop system is initialized at the origin, and the process disturbances and measurement noise are varied at each sampling instance. As illustrated in Fig. 6.7a, based on the values of the monitoring variables at the switching instances, it can be seen that the smart attack is not detected with the scheduled control mode switching approach. Then 1000 simulations of the attacked process with the same process disturbance and measurement noise considered in the simulation with the scheduled switch are performed. Over each simulation, a randomized control mode switching is implemented, where the switching instances are the same as over the corresponding simulations for the linear CSTR process model in Section 6.3.2.1. Over all simulations, detection of the attack occurs within a minimum of 5 time steps (0.05 h in real time) and a maximum of 44 time steps (0.44 h in real time) from the first switching instance. Fig. 6.7b illustrates the evolution of the monitoring variable over a few time steps of the process under the smart attack, with the control mode switch implemented randomly. Over this simulation, attack detection occurs at time step $t_d = 231$ (2.31 h in real time), which is 10 time steps after the first switch from the nominal control mode to the attack-sensitive control mode is implemented at time step $t_{s_1} = 221$ (2.21 h in real time). The reachable set for the attack-free process at the detection time step is contained entirely within the terminal set. The result highlights that, for this simulation, monitoring of the process using a reachable set-based detection scheme may be preferable to monitoring of the process using the terminal set-based detection scheme to aid in the detection of the attack at the earliest time step possible. Following the detection of the attack, the process is switched back to the nominal control mode and no further alarms are observed in the detection scheme.

## 6.4 Conclusions

In this chapter, a cyberattack detection method that utilizes randomized control mode switching to enable the detection of an attack on processes during transient operation was presented. The proposed detection method guarantees no alarms in the detection scheme for the attack-free process with the control mode switching. In developing the detection method, the interdependence between the control parameters, closed-loop stability of the attacked process, and the ability of a reachable set-based detection scheme to detect the attack was exploited. As chemical processes are under prolonged operation at or near their steady states, a modification of the detection method for application to processes operating under steady-state conditions was proposed. Using two illustrative examples, the application of the control mode switching for attack detection with zero false alarms was demonstrated. In the first example, a scalar process under transient operation was considered, while in the second example, a chemical process under steady-state operation was considered. Finally, using simulations of the chemical process example, it was demonstrated that a randomized control mode switch may prevent an attacker from learning the switching schedule, thereby preventing them from designing a smart attack that evades detection.

# Chapter 7

# A Set-Based Control Mode Selection Approach for Active Detection of False Data Injection Cyberattacks

In this chapter, the selection of alternative active detection methods for detecting false-data injection cyberattacks that alter the data communicated over the PCS communication channels is considered. In particular, two alternative control modes, one involving changing set points and the other involving switching control parameters, are considered for the active detection of a class of stealthy false data injection attacks. Implementing either control mode induces perturbations in the closed-loop process. To guarantee the detection of an attack, the perturbations induced from implementing a control mode on the attacked process should be “attack-revealing.” Reachability analysis is used to present a condition that if satisfied means that an attack will be detected, forming the basis of attack-revealing perturbations. Using the condition, a screening algorithm that may be used to choose a control mode that guarantees the detection of an attack is presented. The application of the screening algorithm is demonstrated using an illustrative process example.

# 7.1 Preliminaries

## 7.1.1 Notation

$\mathbb{R}^n$ is the n-dimensional Euclidean space. $\mathbb{Z}^+$ is the set of non-negative integers. For a square matrix $A \in \mathbb{R}^n \times \mathbb{R}^n$, its spectral radius is defined as $\rho(A) = \max_i |\lambda_i(A)|$, where $\lambda_i$ is the $i^{\text{th}}$ eigenvalue of the matrix $A$. Given two sets $\mathcal{X} \in \mathbb{R}^n$ and $\mathcal{Y} \in \mathbb{R}^n$, their Minkowski sum is defined as $\mathcal{X} \oplus \mathcal{Y} = \{x' + y' \mid x' \in \mathcal{X}, y' \in \mathcal{Y}\}$. Given a set $\mathcal{X} \subset \mathbb{R}^n$ and a matrix $A \in \mathbb{R}^n \times \mathbb{R}^n$, the linear map of $\mathcal{X}$ under $A$ is defined as $A\mathcal{X} = \{Ax' \mid x' \in \mathcal{X}\}$, and $\bigoplus_{i=0}^{N} A^i \mathcal{X}$ represents $\mathcal{X} \oplus A\mathcal{X} \oplus \ldots \oplus A^N \mathcal{X}$.

## 7.1.2 Class of Processes

Processes modeled by discrete-time linear time-invariant dynamics are considered:

$$x_{t+1} = A^x x_t + B^u u_t + B^w w_t \tag{7.1a}$$

$$y_t = C^x x_t + v_t \tag{7.1b}$$

where $x_t \in \mathbb{R}^n$ is the state vector, $u_t \in \mathbb{R}^l$ is the manipulated input vector, $w_t \in \mathcal{W} \subset \mathbb{R}^p$ is the process disturbance vector, $y_t \in \mathbb{R}^m$ is the measured output vector, and $v_t \in \mathcal{V} \subset \mathbb{R}^m$ is the measurement noise vector. Without loss of generality, $t = 0$ is taken to be the initial time. The sets $\mathcal{W}$ and $\mathcal{V}$ are convex polytopes. $A^x$, $B^u$, and $C^x$ are matrices of appropriate dimensions, and $B^u$ has full column rank.

A Luenberger observer is used to estimate the states:

$$\hat{x}_{t+1} = A^x \hat{x}_t + B^u u_t + L(y_t - \hat{y}_t) \tag{7.2a}$$

$$\hat{y}_t = C^x \hat{x}_t \tag{7.2b}$$

where $\hat{x}_t \in \mathbb{R}^n$ is the estimated state vector, $\hat{y}_t \in \mathbb{R}^m$ is the estimated output vector, and $L \in \mathbb{R}^n \times \mathbb{R}^m$ is the observer gain. The control objective is to operate at a desired operating steady-state $x^s \in \mathbb{R}^n$. To achieve the control objective, the control input is:

$$u_t = -K(\hat{x} - x^s) + u_s \tag{7.3}$$

where $K \in \mathbb{R}^l \times \mathbb{R}^n$ is the controller gain and $u_s$ is the controller bias used to achieve offset-free control. For simplicity, the expected value of the process disturbance is assumed to be zero. The bias may be computed from:

$$u_s = G^u(I - A^x)x^s \tag{7.4}$$

where $G^u = ((B^u)^T B^u)^{-1}(B^u)^T \in \mathbb{R}^l \times \mathbb{R}^n$ is the left pseudo-inverse of $B^u$. In this work, changing the operating steady-state is considered. The operating steady-state is referred to as the set point for simplicity. The set points are assumed to be selected such that they are reachable in the sense that there exists $u_s \in \mathbb{R}^l$ satisfying Eq. 7.4.

The state estimation error dynamics are given by:

$$e_{t+1} = (A^x - LC^x)e_t + B^w w_t - Lv_t \tag{7.5}$$

where $e_t := x_t - \hat{x}_t \in \mathbb{R}^n$ denotes the state estimation error. The collective dynamics of the closed-loop process encompass both the process states and the estimation error. To facilitate the analysis, an augmented state vector $\xi_t := [x_t^T \ e_t^T]^T$ is defined, and its dynamics are given by:

$$\begin{aligned}\xi_{t+1} = &\underbrace{\begin{bmatrix} A^x - B^uK & B^uK \\ 0 & A^x - LC^x \end{bmatrix}}_{=:A^\xi(K,L)} \xi_t + \underbrace{\begin{bmatrix} B^w & 0 \\ B^w & -L \end{bmatrix}}_{=:B^d(L)} d_t \\ &+ \underbrace{\begin{bmatrix} B^u(G^u(I - A^x) + K) \\ 0 \end{bmatrix}}_{=:B^s(K)} x^s\end{aligned} \tag{7.6}$$

where $d_t := [w_t^T \ v_t^T]^T \in \mathcal{D}$ and $\mathcal{D} := \mathcal{W} \times \mathcal{V}$. For simplicity of presentation, the vector $d_t$ is called the disturbance vector and the set $\mathcal{D}$ is called the disturbance set. Since $\mathcal{W}$ and $\mathcal{V}$ are assumed to be convex polytopes, $\mathcal{D}$ is a convex polytope.

To ensure stable closed-loop behavior, the controller and observer gains are selected such that all eigenvalues of the matrices $A^x - B^uK$ and $A^x - LC^x$ are strictly within the unit circle. Due to the presence of process disturbances and measurement noise, the closed-loop process is persistently perturbed. As a result, the augmented state of the process is ultimately bounded within the minimum invariant set, which is the limit set of all trajectories of the process [79]. The minimum invariant set of the process is given by [77]:

$$\mathcal{R}^{\xi}_{\infty}(x^s) = \bigoplus_{i=0}^{\infty} A^{\xi}(K, L)^i \mathcal{D}^e(x^s) \tag{7.7}$$

where $\mathcal{D}^e(x^s) = B^d(L)\mathcal{D} \oplus B^s(K)\{x^s\}$.

### 7.1.3 Class of False Data Injection Attacks

The process is vulnerable to false data injection (FDI) attacks. These attacks alter the output ($y_t$) transmitted via the sensor-controller link and the input ($u_t$) conveyed over the controller-actuator link so that the altered values are received by the controller and actuators. Additive and multiplicative FDI attacks are considered where the relationships between the unaltered and altered values may be described as:

$$y^a_t = \Lambda^y y_t + \delta^y_t \tag{7.8a}$$

$$u^a_t = \Lambda^u u_t + \delta^u_t \tag{7.8b}$$

If $\Lambda^\theta \neq I$ ($\theta \in \{y, u\}$), the attack alters the data over a communication link by multiplying it with the factor $\Lambda^\theta$. If $\delta^\theta_t \neq 0$, the attack alters the data over a communication link by adding a bias $\delta^\theta_t$. $\theta = y$ represents the sensor-controller link, and $\theta = u$ represents the controller-actuator link. The variables $\delta^y_t$ and $\delta^u_t$ are assumed to be bounded within a convex polytope, i.e., $\delta_t := \begin{bmatrix} \delta^u_t \\ \delta^y_t \end{bmatrix} \in \Delta$ for all $t \in \mathbb{Z}^+$

An attack on the closed-loop process alters the evolution of its augmented state as follows:

$$\begin{aligned}\xi_{t+1} = {} & \underbrace{\begin{bmatrix} A^x - B^u\Lambda^u K & B^u\Lambda^u K \\ L(I-\Lambda^y)C^x & A^x - LC^x \end{bmatrix}}_{=:A^{\xi_a}(K,L)} \xi_t + \underbrace{\begin{bmatrix} 0 & B^u \\ -L & 0 \end{bmatrix}}_{=:B^{\delta_a}(L)} \delta_t \\ & + \underbrace{\begin{bmatrix} B^w & 0 \\ B^w & -L\Lambda^y \end{bmatrix}}_{=:B^{d_a}(L)} d_t + \underbrace{\begin{bmatrix} B^u\Lambda^u(G^u(I-A^x)+K) \\ 0 \end{bmatrix}}_{=:B^{s_a}(K)} x^s \end{aligned} \tag{7.9}$$

Eq. 7.9 is formulated with the assumption that the observer states are driven by the implemented control action ($u^a_t$). From Eq. 7.9, the process can be destabilized by a multiplicative attack, i.e., $\rho(A^{\xi_a}(K,L)) > 1$. However, an additive attack with $\Lambda^u = I$ and $\Lambda^y = I$ does not alter the closed-loop stability.

When the process is subjected to an FDI attack, the process is referred to as the attacked process. The term attack-free process is used to describe the closed-loop process without an attack. When the attacked process is stable, the augmented states are ultimately bounded within its minimum invariant set, which is a compact set. The minimum invariant set is:

$$\mathcal{R}^{\xi}_{\infty}(x^s) = \bigoplus_{i=0}^{\infty} A^{\xi_a}(K,L)^i \mathcal{D}^a(x^s) \tag{7.10}$$

where $\mathcal{D}^a(x^s) := B^{d_a}\mathcal{D} \oplus B^{\delta_a}\Delta \oplus B^{s_a}(K)\{x^s\}$. If the attack is destabilizing, the set $\mathcal{R}^{\xi}_{\infty}(x^s)$ is unbounded.

### 7.1.4 Monitoring Variable and Set-Based Detection Scheme

An attack alters the evolution of the augmented state from its expected attack-free evolution. However, since the augmented state cannot be measured directly, detection schemes monitor the evolution of a monitoring variable to detect anomalous behavior. A monitoring variable ($\eta := [y^T \; r^T]^T$) that is a concatenation of the measured output and the residual vector ($r := y - \hat{y}$) is used to monitor the process. For the attack-free process, its monitoring variable may be expressed as a linear combination of the augmented state and the disturbance vectors:

$$\eta_t = \underbrace{\begin{bmatrix} C^x & 0 \\ 0 & C^x \end{bmatrix}}_{=:C^{\eta}} \xi_t + \underbrace{\begin{bmatrix} 0 & I \\ 0 & I \end{bmatrix}}_{=:D^{\eta}} d_t \tag{7.11}$$

For the attacked process, its monitoring variable may be expressed as a linear combination of the augmented state, the disturbance vector, and the attack biases added to the sensor-controller link as:

$$\eta_t = \underbrace{\begin{bmatrix} \Lambda^y C^x & 0 \\ (\Lambda^y - I)C^x & C^x \end{bmatrix}}_{=:C^{\eta_a}} \xi_t + \underbrace{\begin{bmatrix} 0 & \Lambda^y \\ 0 & \Lambda^y \end{bmatrix} d_t + \begin{bmatrix} \delta^y_t \\ \delta^y_t \end{bmatrix}}_{=:d^a_t} \tag{7.12}$$

where $d^a_t \in \mathcal{D}^{\eta_a}(x^s) := \begin{bmatrix} 0 & \Lambda^y \\ 0 & \Lambda^y \end{bmatrix} \mathcal{D} \oplus \begin{bmatrix} 0 & I \\ 0 & I \end{bmatrix} \Delta$ for all $t \in \mathbb{Z}^+$.

Chemical processes typically operate at steady-state for extended duration, ensuring that the augmented state remains within its minimum invariant set. In the absence of attacks, the monitoring variable is contained within a well-defined set when $\xi_t \in \mathcal{R}^{\xi}_{\infty}(x^s)$. From Eq. 7.10 and Eq. 7.11, this set, called the terminal set, is represented as:

$$\mathcal{R}^{\eta}_{\infty}(\mathcal{R}^{\xi}_{\infty}(x^s)) = C^{\eta}\mathcal{R}^{\xi}_{\infty}(x^s) \oplus D^{\eta}\mathcal{D} \tag{7.13}$$

The terminal set for the attack-free process encompasses all conceivable values of the monitoring variable across all time steps ($t \in \mathbb{Z}^+$) and under all disturbances ($d_t \in \mathcal{D}$) when $\xi_t \in \mathcal{R}^{\xi}_{\infty}(x^s)$. Therefore, the terminal set can be used to verify the integrity of monitoring variable values. To monitor for attacks, a terminal set membership-based detection scheme is employed:

$$\phi(\eta_t) = \begin{cases} 0, \ \eta_t \in \mathcal{R}^{\eta}_{\infty}(\mathcal{R}^{\xi}_{\infty}(x^s)) \\ 1, \ \eta_t \notin \mathcal{R}^{\eta}_{\infty}(\mathcal{R}^{\xi}_{\infty}(x^s)) \end{cases} \tag{7.14}$$

where $\phi : \mathbb{R}^{2m} \to \{0, 1\}$ is the detection scheme mapping. The scheme generates an output of 0 when the monitoring variable resides within its attack-free terminal set, meaning that no attack is detected. If the monitoring variable falls outside the attack-free terminal set, the detection scheme outputs a 1, indicating the detection of an attack.

If the process operates for a sufficiently long period after an attack and the closed-loop process is stable, the augmented state will converge to the minimum invariant set under the attack ($\mathcal{R}^{\xi_a}_{\infty}(x^s)$). For analysis purposes, the corresponding terminal set of the monitoring variable can be computed from $\mathcal{R}^{\xi_a}_{\infty}(x^s)$, given by:

$$\mathcal{R}^{\eta_a}_{\infty}(\mathcal{R}^{\xi_a}_{\infty}(x^s)) = C^{\eta_a}\mathcal{R}^{\xi_a}_{\infty}(x^s) \oplus \mathcal{D}^{\eta_a}(x^s) \tag{7.15}$$

## 7.2 Active Detection for Stealthy Attacks

The terminal set-based detection scheme in Eq. 7.14 is passive, as it monitors the process for attacks without utilizing any external intervention or perturbations. An attacker may be able to carry out stealthy attacks that are capable of evading detection. Stealthy attacks with respect to the passive terminal set-based detection scheme in Eq. 7.14 may be defined as attacks such that the monitoring variable of the attacked process is such that $\eta_t \in \mathcal{R}_\infty^\eta(\mathcal{R}_\infty^\xi(x^s))$ for all time steps $t \in \mathbb{Z}^+$. As a result, the terminal set-based detection scheme generates an output of 0 and fails to detect a stealthy attack.

An active detection method utilizing an external intervention to perturb the attacked process may be used to enable the detection of a stealthy attack. In prior chapters and work, an active detection method that utilized changing control parameters $(K, L)$ online to enhance the detection capabilities of multiplicative FDI attacks was presented [48, 50]. Specifically, the control system intermittently switches to operate in the so-called attack-sensitive mode. The control parameters for the attack-sensitive mode are selected such that the attack-free process is stable, but some attacks will destabilize the process. Under certain conditions, these attacks will be detected under the attack-sensitive mode.

The active detection method involving operating under an attack-sensitive mode could be applied to detect stealthy attacks. However, operating in a mode that allows some attacks to destabilize the process may be undesirable. For example, an unstable process can result in exponential growth in the states, which may lead to the states breaching the safety limits of process equipment. Moreover, attack-sensitive control parameters may not exist for certain attacks. For example, an additive attack, i.e., an attack such that $\Lambda^u = I$ and $\Lambda^y = I$, does not destabilize the process. Therefore, alternative active detection methods should be considered. In the literature, several active detection methods have been proposed (e.g., [104–111]). In this work, two active detection methods are considered: changing the control parameters and the set point, which define alternative operating modes of the control system to enable the detection of stealthy attacks. A framework for evaluating if an attack is guaranteed to be detected under this active method is developed for these control modes.

### 7.2.1 Active Detection with Set Point and Control Parameter Changes

The process after extended operation near the initial set point $x_i^s$ under control parameters $(K^i, L^i)$ is considered so that the augmented state has converged to either $\mathcal{R}_\infty^\xi(x_i^s)$ or $\mathcal{R}_\infty^{\xi_a}(x_i^s)$, depending on whether the process is attack-free or subjected to an attack. Under the active detection method, a set point change from $x^s = x_i^s$ to $x^s = x_f^s$ and/or a control parameter switch from $(K, L) = (K^i, L^i)$ to $(K, L) = (K^f, L^f)$ is implemented. Without loss of generality, the time step in which the active detection method is implemented at $t = 0$. While not essential to the problem formulation, the initial set point and initial control parameters $(K^i, L^i)$ are taken to be selected based on the desired operating set point and standard controller design methods. However, the final set point and/or final control parameters are selected to guarantee attack detection. Ultimately, the goal is to select the set point and control parameters to enable the detection of a given attack, defined by $\Lambda^u$, $\Lambda^y$, and $\Delta$, that is stealthy with respect to the terminal set-based detection scheme. The attack-free and attacked process are stable under both sets of control parameters in the sense that $\rho(A^{\xi_a}(K^i, L^i)) < 1$ and $\rho(A^{\xi_a}(K^f, L^f)) < 1$ (instability under an attack as a mechanism for attack detection has been previously addressed [48, 50]).

A set point change and/or control parameter change perturbs the process by exciting the process dynamics. From Eq. 7.11 and Eq. 7.12, the values of the monitoring variable depend upon the augmented state, the disturbances acting on the process, and the attack (in the case of the attacked process). For the closed-loop stable process, a perturbation causes the process states to evolve outside the minimum invariant set under the initial steady-state for a transient period until the states are ultimately bounded within the minimum invariant set of the process at the final steady-state. During the transient period, the monitoring variable values may not be contained within their terminal sets. The terminal set-based detection scheme in Eq. 7.14 does not account for transient behavior in the attack-free process, and generates false alarms during the transient period after implementation of the active detection method. To monitor the process during the transient period with no false alarms, a reachable set-based detection scheme may be used [49].

During the transient period, the possible states reached for the attack-free and attacked process are described by the reachable sets of the process [49]. From Eq. 7.6 and Eq. 7.9, the reachable sets of the attack-free and the attacked process are:

$$\mathcal{R}_t^{\xi}(x_f^s) = A^{\xi}(K, L)\mathcal{R}_{t-1}^{\xi}(x_f^s) \oplus \mathcal{D}^e(x_f^s) \tag{7.16a}$$

$$\mathcal{R}_t^{\xi_a}(x_f^s) = A^{\xi_a}(K, L)\mathcal{R}_{t-1}^{\xi_a}(x_f^s) \oplus \mathcal{D}^a(x_f^s) \tag{7.16b}$$

for $t > 0$ where, with slight abuse of notation, the initial sets are $\mathcal{R}_0^{\xi} = \mathcal{R}_{\infty}^{\xi}(x_i^s)$ and $\mathcal{R}_0^{\xi_a} = \mathcal{R}_{\infty}^{\xi_a}(x_i^s)$. Eq. 7.16a describes the evolution of the reachable sets for the attack-free augmented states from an initial set of states that is the minimum invariant set of the attack-free process at the initial steady-state. Eq. 7.16b describes the evolution of the reachable sets for the augmented states of the attacked process from an initial set of states that is the minimum invariant set of the attacked process at the initial steady-state. From Eq. 7.11 and Eq. 7.12, the reachable sets of the monitoring variables describe their evolution for the attacked and the attack-free processes as:

$$\mathcal{R}_t^{\eta}(\mathcal{R}_t^{\xi}(x_f^s)) = C^{\eta}\mathcal{R}_t^{\xi}(x_f^s) \oplus D^{\eta}\mathcal{D} \tag{7.17a}$$

$$\mathcal{R}_t^{\eta_a}(\mathcal{R}_t^{\xi_a}(x_f^s)) = C^{\eta_a}\mathcal{R}_t^{\xi_a}(x_f^s) \oplus \mathcal{D}^{\eta_a} \tag{7.17b}$$

for $t > 0$.

For the attack-free process, its monitoring variable values evolve within its reachable sets. A reachable set-based detection scheme designed to utilize the reachable sets for the attack-free process as a certificate to verify the fidelity of the monitoring variable values is used to monitor the process for attacks [49]:

$$\phi_t(\eta_t) = \begin{cases} 0, \ \eta_t \in \mathcal{R}_t^{\eta}(\mathcal{R}_t^{\xi}(x_f^s)) \\ 1, \ \eta_t \notin \mathcal{R}_t^{\eta}(\mathcal{R}_t^{\xi}(x_f^s)) \end{cases} \tag{7.18}$$

where $\phi_t(\eta_t)$ is the output of the detection scheme at the time step $t > 0$. The detection scheme generates an output of 1 if the monitoring variable is not contained within its attack-free reachable set, meaning that an attack is detected. However, if the monitoring variable is contained within the attack-free reachable set, then the detection scheme generates an output of 0 indicating a lack of attack detection.

### 7.2.2 Selecting an Attack-Revealing Control Mode for Active Detection

From Eq. 7.17a and Eq. 7.17b, the monitoring variable values for the attack-free and the attacked processes are contained within their respective reachable sets. If at some time, the reachable sets of the monitoring variable for the attacked and the attack-free processes at that time step do not intersect, then there exist no values of monitoring variable values of the attacked process, that are contained within the reachable set of the attack-free process. It follows from this reasoning that the perturbation induced by switching the control mode is attack-revealing if, at some time step $t > 0$, the reachable sets of the monitoring variable for the attacked and the attack-free process satisfy:

$$\mathcal{R}_t^{\eta}(\mathcal{R}_t^{\xi}(x_f^s)) \cap \mathcal{R}_t^{\eta_a}(\mathcal{R}_t^{\xi_a}(x_f^s)) = \emptyset \tag{7.19}$$

The reachable set-based detection scheme in Eq. 7.18 monitors the process based on the reachable sets for the attack-free process, attack detection is guaranteed at the time step $t$ if the perturbation induced is attack-revealing, i.e., if Eq. 7.19 is satisfied.

In the discussion that follows, a screening algorithm that leverages Eq. 7.19 to enable the selection of a control mode that guarantees attack detection is presented. The algorithm is implemented offline and requires that the reachable sets for the attacked and the attack-free processes operated under a given control mode be computed, and the satisfaction of Eq. 7.19 be checked. If at some time step $t_d \in \mathbb{Z}^+$, the reachable sets of the attack-free and the attacked processes satisfy Eq. 7.19, then the control mode chosen induces attack-revealing perturbations in that it guarantees the attack detection. However, if Eq. 7.19 is never satisfied, then it means that the control mode induced does not guarantee attack detection.

Before introducing the algorithm, a practical implementation challenge is discussed. Ensuring the satisfaction of Eq. 7.19 requires computing reachable sets for both the attack-free and attacked processes, potentially extending to an infinite number of time steps, which is infeasible. The algorithm must strike a balance between checking a finite number of reachable sets and the computational complexity. More specifically, a possibility exists that the condition in Eq. 7.19 is satisfied for some time step after the time step that the algorithm is terminated. To manage this tradeoff, a parameter $t_f > 0$ is introduced, which is the number of time steps to compute the reachable sets before terminating the algorithm. Opting for a large $t_f$ may reduce the possibility that Eq. 7.19 is satisfied for some time after $t_f$ but may increase computational demands, while selecting a small $t_f$ may heighten this possibility but may reduce computation. This is grounded in the understanding that, given an error threshold, there exists a time duration large enough for the reachable sets to converge to an invariant set containing the minimum invariant set. The discrepancy between the invariant set and the true minimum invariant set depends on the chosen error threshold [77, Theorem 1]. On a more practical level, choosing $t_f$ could involve selecting the number of time steps at which it becomes essential to detect the attack, especially if operating under the alternative control mode for prolonged periods is undesirable.

The inputs to the screening algorithm include the attack model ($\Lambda^y$, $\Lambda^u$, and $\Delta$), the reachable sets of the augmented state, the final steady-state $x^s_f$, the control parameters $(K^f, L^f)$, and the termination time step $t_f$. The algorithm is initialized at the time step $t = 0$, where the initial sets for the attacked and the attack-free processes are their respective minimum invariant sets at the initial steady-state. The main part of the algorithm involves checking if Eq. 7.19 is satisfied. The screening algorithm is terminated if Eq. 7.19 is satisfied at $t_d < t_f$, indicating that attack detection is guaranteed at $t_d$ or if Eq. 7.19 is not satisfied at the time step $t = t_f$, indicating that attack detection is not guaranteed.

**Remark 7.2.1.** *A systematic approach to determining $t_f$ involves tracking the convergence of reachable sets to their respective minimum invariant sets. More precisely, as outlined in [79], invariant outer $\epsilon$-approximations of the minimum invariant sets can be calculated by incorporating a predefined error threshold ($\epsilon > 0$). By computing the reachable sets for both the attack-free and attacked processes and verifying their containment within the approximate minimum invariant sets, the convergence of the reachable sets can be ascertained. If both reachable sets converge to their minimum invariant sets, encapsulating them completely, two times, $t_1$ and $t_2$, are defined, representing when the attack-free and attacked process reachable sets converge to the approximate minimum invariant set. In this case, the choice of termination time step is $t_f = \max(t_1, t_2)$. This method offers control over the error caused by limiting the computation to a finite number of reachable sets.*

**Algorithm 4:** Algorithm to screen an active detection method for its ability to guarantee attack detection

```
Inputs: $\Lambda^y$, $\Lambda^u$, $\Delta$, $(K^f, L^f)$, $x_f^s$, $t_f$, $\mathcal{R}_t^{\xi}(x_f^s)$ and $\mathcal{R}_t^{\xi_a}(x_f^s)$ for $t \in (0, t_f]$.
Initialization: $t = 0$, $\mathcal{R}_0^{\xi} = \mathcal{R}_{\infty}^{\xi}(x_i^s)$, $\mathcal{R}_0^{\xi_a} = \mathcal{R}_{\infty}^{\xi_a}(x_i^s)$, $t_d = \infty$, $(K, L) = (K^f, L^f)$
do
    Compute the reachable sets for the monitoring variable in Eq. 7.17a and Eq. 7.17b.
    if Eq. 7.19 is satisfied then
        The chosen control mode guarantees attack detection at $t_d = t$.
    else if $t = t_f$ then
        The chosen control mode does not guarantee attack detection.
    else
        Set $t \leftarrow t + 1$
while $t_d = \infty$;
```

## 7.3 Application to an Illustrative Process

A process under a simultaneous sensor-controller link and controller-actuator link FDI attack is considered:

$$
\begin{aligned}
x_{t+1} &= x_t + u_t^a + w_t \\
u_t^a &= -\Lambda^u K(\hat{x}_t - x^s) + \delta_t^u \\
y_t^a &= \Lambda^y (x_t + v_t) + \delta_t^y
\end{aligned}
$$

where $x_t \in \mathbb{R}$ is the state, $u_t^a \in \mathbb{R}$ is the control action received by the control actuators, $w_t \in \mathcal{W} := \{w' \mid |w'| \leq 1\}$ is the process disturbance, $y_t^a \in \mathbb{R}^m$ is the measured output received by the controller, and $v_t \in \mathcal{V} := \{v' \mid |v'| \leq 1\}$ is the measurement noise. For this integrating process, there is an equilibrium manifold corresponding to the steady-state input $u^s = 0$. The initial steady-state is the origin, i.e., $x_i^s = 0$. The control parameters chosen to operate the process at the initial steady-state are $(K^i, L^i) = (0.8541, 0.618)$. An attack with $\Lambda^y = 0.86$, $\Lambda^u = 1.1$, $\delta_t^y = 0.1$, and $\delta_t^u = -0.028$ is considered. The MPT 3.0 toolbox is used for polytope computations [82].

The following results encompass the evaluation of various attack detection methods and schemes. To verify the detectability characteristics of the attack across these methods and schemes, 1000 simulation scenarios have been generated. Within each scenario, an initial condition is randomly selected from the minimum invariant set of the attacked process ($\mathcal{R}_{\infty}^{\xi_a}(x_i^s)$). Additionally, random sequences are generated to simulate process disturbances and measurement noise. Each element within these sequences is drawn from $\mathcal{N}(0, 3.33 \times 10^{-2})$ and each simulation scenario spans 100 time steps.

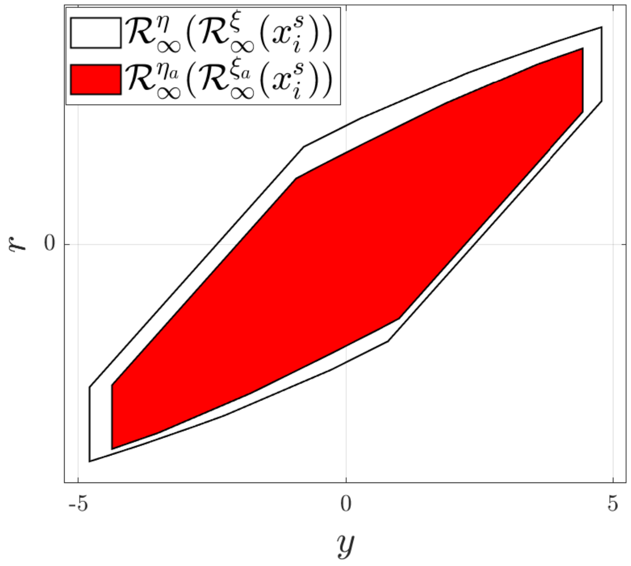


Fig. 7.1: Terminal sets of the monitoring variable for the attacked and the attack-free processes at the initial steady-state $x_i^s = 0$ with control parameters $(K^i, L^i)$.

The detectability properties of this attack under the terminal set-based detection scheme in Eq. 7.14 are first investigated. To analyze the detectability of the attack under the detection scheme, the terminal sets of the monitoring variable for the attack-free and the attacked process are computed. The terminal set of the attacked process is contained entirely within the terminal set of the attack-free process, i.e., $\mathcal{R}_{\infty}^{\eta_a}(\mathcal{R}_{\infty}^{\xi_a}(x_i^s)) \subset \mathcal{R}_{\infty}^{\eta}(\mathcal{R}_{\infty}^{\xi}(x_i^s))$ (Fig. 7.1). This implies that the attack is undetectable, i.e., stealthy with respect to the detection scheme because for any $\xi_t \in \mathcal{R}_{\infty}^{\xi_a}(x_i^s)$, $\eta_t \in \mathcal{R}_{\infty}^{\eta_a}(\mathcal{R}_{\infty}^{\xi_a}(x_i^s)) \subset \mathcal{R}_{\infty}^{\eta}(\mathcal{R}_{\infty}^{\xi}(x_i^s))$. To verify the undetectability of the attack, 1000 closed-loop simulations of the attacked process are performed. The attack is not detected in any of these simulations. To enable the detection, an active detection method may be needed to detect this attack.

One active detection option is to change the control parameters to so-called attack-sensitive parameters, where the closed-loop process with these parameters is stable under attack-free operation and is destabilized under the attack. One choice for control parameters for the attack-sensitive mode is $K = 1.57$ and $L = 1.28$. To analyze the detection under this active detection method, 1000 simulations of the attack process are performed where the control parameters change from the initial control parameters to the attack-sensitive parameters at the initial time. A reachable set-based detection scheme (Eq. 7.18) is utilized to monitor the process. The attack is detected over 996 simulations. For the simulations where the attack is detected, the detection times ranged from 28 to 96 times. These results demonstrate that operating under the attack-sensitive mode enhances the detection capabilities. However, under the attack-sensitive mode, the attack renders the closed-loop process unstable, which may be undesirable, and alternative active detection methods may be preferable.

Other active detection methods that may enable attack detection without destabilization are considered. To check if a particular active detection method guarantees attack detection, Algorithm 4 is applied and implemented as follows. The method described in [79] is used to compute outer approximation of the minimum invariant sets of the attack-free and attacked processes for the initial steady-state and control parameters and the final steady-state and control parameters, i.e., the sets $\mathcal{R}_\infty^{\xi}(x_i^s)$, $\mathcal{R}_\infty^{\xi_a}(x_i^s)$, $\mathcal{R}_\infty^{\xi}(x_f^s)$, and $\mathcal{R}_\infty^{\xi_a}(x_f^s)$. The specified error bound on these calculations is $1 \times 10^{-3}$. To determine the termination time of the algorithm ($t_f$), the reachable sets of the attack and attack-free process with initial sets $\mathcal{R}_\infty^{\xi}(x_i^s)$ and $\mathcal{R}_\infty^{\xi_a}(x_i^s)$ are computed until the reachable sets are contained within the outer approximation of the attack-free and attacked minimum invariant sets for the final steady-state and control parameters. Defining $t_1$ and $t_2$ as the time steps at which the attack-free and attacked reachable sets are first contained with their corresponding minimum invariant sets, respectively, $t_f$ is taken to be $\max(t_1, t_2)$. The satisfaction of Eq. 7.19 is verified by checking for the existence of a point satisfying both sets of inequalities describing the two reachable sets of the monitoring variable for the attack-free and the attacked processes. Specifically, a feasibility problem, cast as a linear program, is constructed and solved for all $t \in [0, t_f]$.

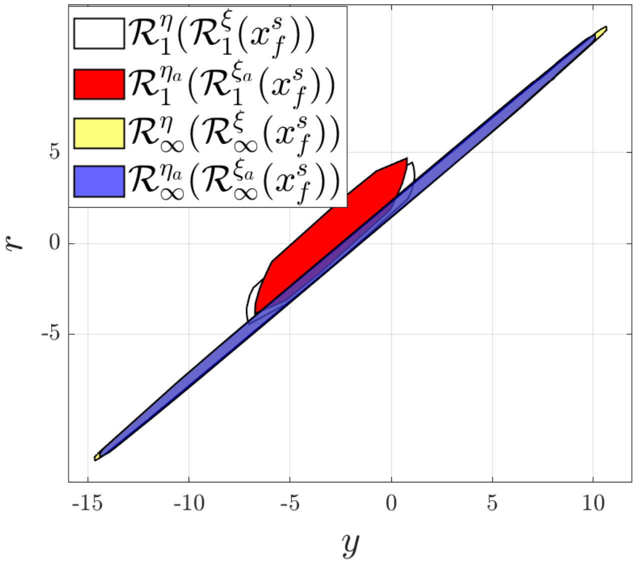


Fig. 7.2: Reachable sets of the monitoring variable at $t = 1$ and the terminal sets of the monitoring variable for the attacked and the attack-free process under a control mode with $x_f^s = -2$ and $(K^f, L^f) = (1.5, 0.1)$.

The first alternative active detection method considered utilizes a set point change to shift the operation of the process to a neighborhood of the steady-state $x_f^s = -2$ and with the control parameters $(K^f, L^f) = (1.5, 0.1)$. Under these control parameters, the eigenvalues of the closed-loop attacked process are -0.67 and 0.92, indicating that the closed-loop process is stable in the presence of attack. In this case, $t_1 = 53$ and $t_2 = 75$, so $t_f = 75$. Using Algorithm 4, Eq. 7.19 is not satisfied for any time step, and therefore, attack detection is not guaranteed. Fig. 7.2 illustrates the reachable sets of monitoring variable for the attacked and the attack-free processes at the time step $t = 1$, and the terminal sets of the monitoring variable for the attacked and the attack-free processes under the chosen active detection method, showing that Eq. 7.19 is not satisfied. From Fig. 7.2, parts of both sets $\mathcal{R}_1^{\eta_a}(\mathcal{R}_1^{\xi_a}(x_f^s))$ and $\mathcal{R}_\infty^{\eta_a}(\mathcal{R}_\infty^{\xi_a}(x_f^s))$ are not contained within the attack-free sets $\mathcal{R}_1^{\eta}(\mathcal{R}_1^{\xi}(x_f^s))$ and $\mathcal{R}_\infty^{\eta}(\mathcal{R}_\infty^{\xi}(x_f^s))$, respectively, indicating that attack detection is (theoretically) possible. However, the lack of separation of these sets, indicates that attack detection is not guaranteed.

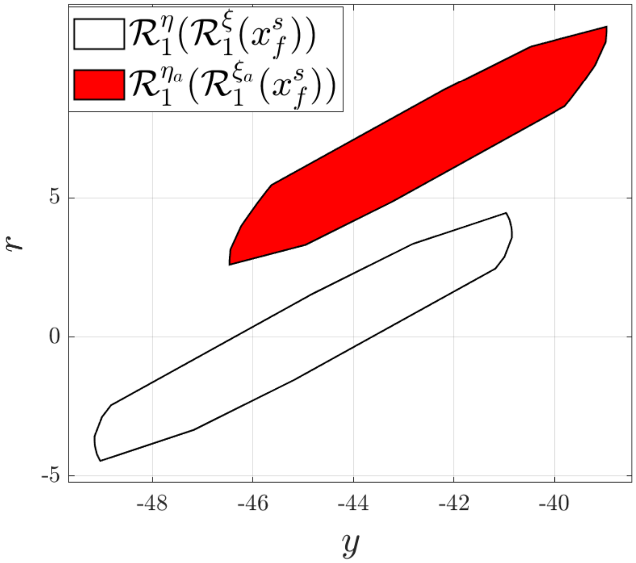


Fig. 7.3: Reachable sets of the monitoring variable for the attacked and the attack-free process at $t_d = 1$ under a control mode with $x_f^s = -30$ and $(K^f, L^f) = (1.5, 0.1)$.

To verify the detection properties of the attack under the first alternative active detection method, 1000 closed-loop simulations of the attacked process monitored by the reachable set-based detection scheme in Eq. 7.18 are performed under the active detection method. The attack is detected in 114 simulations. For the simulations where the attack is detected, the attack is detected at either time step 1 or 2. While the attack may be detected with this active detection method, detection is not guaranteed.

A second active detection method using a set point change to $x_f^s = -30$, and a control parameter switch to $(K^f, L^f) = (1.5, 0.1)$ is considered. The reachable sets of the augmented state for the attacked and the attack-free processes are computed and the termination time step for Algorithm 4 is determined to be as $t_f = \max(t_1, t_2) = 119$, with $t_1 = 92$ and $t_2 = 119$. Algorithm 4 is applied, and the control mode chosen is determined to guarantee attack detection at the time step $t_d = 1$. Fig. 7.3 illustrates that the reachable sets for the attacked and the attack-free processes do not intersect at the time step $t_d = 1$, satisfying Eq. 7.19. One thousand simulations of the attacked process under this active detection method and monitored by the reachable set-based detection scheme are performed. The attack is detected in all simulations at the time step $t_d = 1$, demonstrating that the active detection method chosen guarantees attack detection. This active detection method guarantees detection, detects the attack quicker than the other methods, and does not result in an unstable closed-loop process under the attack.

## 7.4 Conclusions

Two control modes for the active detection of a class of stealthy false data injection cyber-attacks were presented. Reachability analysis was used to present a screening algorithm that may be used to select an active detection method that guarantees attack detection. The application of the screening algorithm was demonstrated using an illustrative process example.

# Chapter 8

# Conclusions

In this dissertation, approaches for control-enabled active detection of cyberattacks on process control systems (PCSs) were presented. Approaches presented considered the detection of false data injection (FDI) attacks, which alter the operational data over PCS communication links. First, a characterization is performed of the influence of control parameters on the ability of a class of passive detection schemes (that aid in attack detection without external intervention) to detect attacks which is defined as attack detectability. The characterization is exploited to present a controller screening algorithm that may be used to identify and discard control parameters that may mask an attack from the detection scheme. Even when the control parameters are chosen so that they do not mask an attack, the attack may be designed to evade detection by a passive detection this. Realizing this, an active detection method that enhances the attack detection capabilities of the passive detection scheme using an external intervention in the form of a PCS design parameter switch was proposed. PCS design parameter switching on a process under steady-state operation may excite process dynamics and trigger false alarms in the detection scheme. An active detection scheme that schedules PCS parameter switching to avoid transient behavior was proposed for false alarm minimization. To account for transient process behavior, a reachable set-based attack detection scheme was proposed. An active detection method utilizes the reachable set-based attack detection scheme, and randomized PCS design parameter switching was proposed to enable attack detection, with zero false alarms, and while guaranteeing a zero false alarm rate. The PCS design parameter switching active detection methods proposed enabled attack detection by utilizing a switch to PCS design parameter chosen such that an attack destabilizes the process. Destabilizing for attack detection may not be preferred. Therefore, an alternative attack detection method utilizing a PCSs design parameter switching and/or setpoint change was presented for attack detection without destabilization.

Work presented in this dissertation considered FDI attacks that may be modeled as multiplicative and/or additive attacks. Future work may focus on the development of methods for control-enabled detection of other FDI attacks (e.g., replay attacks) may be a possible subject for extending the present work. Development of approaches for alternative control mode selection based on the attack model may also be explored. Finally, all theoretical work presented in this dissertation have considered processes modeled by LTI dynamics. Therefore, extension of results presented to nonlinear processes modeled may be explored.

# Appendix A

# Proofs for Chapter 2

**Theorem 3.** *Consider the closed-loop process represented by the dynamics in Eq. 2.1 under a multiplicative sensor-controller link attack of magnitude $\Lambda \neq I$ with the controller in Eq. 2.4 using the state estimate from the observer in Eq. 2.3 and monitored by a detection scheme that fits the model for the class of residual-based detection scheme in Eq. 2.11. Let the closed-loop process be stable in the sense that all the eigenvalues of $A_\xi$ and $A_{\xi_a}$ (Eq. 2.6) are within the unit circle. If $D_r^{est}$ and $D_{r_a}^{est}$ are numerical estimates of residual sets computed based on Eq. 2.16 and $R_a^{est} \leq R_e^{est}$ where $R_e^{est} := \max\limits_{r' \in D_{r_e}^{est}} \|r'\|$ and $D_{r_e}^{est} := D_r^{est} \ominus A_r B_\infty^n(\epsilon)$, then the attack is undetectable.*

*Proof.* From Eq. 2.17, $D_r^{est} \subseteq D_r \oplus A_r B_\infty^n(\epsilon)$ so, by the standard property of Minkowski difference of sets[112]

$$D_{r_e}^{est} := D_r^{est} \ominus (A_r B_\infty^n(\epsilon)) \subseteq D_r \tag{A.1}$$

making $D_{r_e}^{est} \subseteq D_r$ an inner approximation of $D_r$. Since $D_{r_e}^{est}$ is an inner approximation of $D_r$, $R_e^{est} \leq R$ where $R_e^{est} := \max\limits_{r' \in D_{r_e}^{est}} \|r'\|$. The set $D_{r_a}^{est}$ is an outer approximation of $D_{r_a}$ and $R_a^{est} \geq R_a$. Therefore, $R_a^{est} \leq R_e^{est}$ implies that:

$$R_a \leq R_a^{est} \leq R_e^{est} \leq R$$

or $R_a \leq R$, and the attack is undetectable. □

# Appendix B

# Proofs for Chapter 3

**Proposition 13.** *Consider the closed-loop process operated at steady-state with control system parameters* $(K, L)$ *under a multiplicative sensor-controller link attack of magnitude* $\Lambda$*. If the attack is such that the closed-loop process remains stable, i.e., the eigenvalues of* $A_\xi(\Lambda, K, L)$ *lie within the unit circle, the multiplicative attack is undetectable with respect to the detection scheme in Eq. 3.10, if and only if* $D_r(\Lambda, K, L) \subseteq D_r(I, K, L)$*.*

*Proof.* Consider the closed-loop process operated at steady-state with control system parameters $(K, L)$ under a multiplicative sensor-controller link attack of magnitude $\Lambda$. If the pair $(K, L)$ are stabilizing under the attack, then the minimum invariant set of the process $D_\xi(\Lambda, K, L)$ is compact and forward invariant. Additionally, the augmented state of the attacked closed-loop process  is bounded within its minimum invariant set for all time, i.e., $\xi(t) \in D_\xi(\Lambda, K, L)$ for all $t \geq 0$. As a result, the residuals of the attacked closed-loop process  are also bounded within the terminal set of residuals, i.e., $r(t) \in D_r(\Lambda, K, L)$ for all $\xi(0) \in D_\xi(\Lambda, K, L)$ and $\mathbf{d} \in \mathcal{F}$. If the terminal residual set of the attacked process is a subset of or equal to the terminal residual set of the attack-free process ($D_r(\Lambda, K, L) \subseteq D_r(I, K, L)$), the residuals of the attack process are contained within its attack-free terminal residual set, i.e., $r(t) \in D_r(\Lambda, K, L) \subseteq D_r(I, K, L)$ for all $t \geq 0$ and the attack is undetectable. Hence, the attack is undetectable if $D_r(\Lambda, K, L) \subseteq D_r(I, K, L)$. To show that $D_r(\Lambda, K, L) \subseteq D_r(I, K, L)$ is also a necessary condition for undetectability, the proof proceeds by contradiction. Assume there is an undetectable multiplicative sensor-controller link attack of magnitude $\Lambda$ on the closed-loop process with control system parameters $(K, L)$ such that the attacked closed-loop process is stable and the terminal residual set of the attacked process is not a subset of or equal to the terminal residual set of the attack-free process, i.e., $D_r(\Lambda, K, L) \nsubseteq D_r(I, K, L)$. Based on the definition of undetectable attacks, the multiplicative sensor-controller link attack is such that for any augmented state initialized in the minimum invariant set of the attacked process $\xi(0) \in D_\xi(\Lambda, K, L)$ and all $\mathbf{d} \in \mathcal{F}$, the residuals of the attacked process are contained within the attack-free terminal residual set, i.e., $r(t) \in D_r(I, K, L)$ for all time $t \geq 0$. Since $D_r(\Lambda, K, L) \nsubseteq D_r(I, K, L)$, the set $D_r(\Lambda, K, L) \setminus D_r(I, K, L)$ is non-empty. Moreover, there exist $\xi(0) \in D_\xi(\Lambda, K, L)$ and $\mathbf{d} \in \mathcal{F}$ that result in $r(t) \in D_r(\Lambda, K, L) \setminus D_r(I, K, L)$ for some $t \geq 0$ implying that $r(t) \notin D_r(I, K, L)$ for some $t \geq 0$. This leads to a contradiction, completing the proof. □

**Proposition 14.** *Consider the closed-loop process operated at steady-state with control system parameters* $(K, L)$ *under a multiplicative sensor-controller link attack of magnitude* $\Lambda$*. If the attack is such that (1) the attacked closed-loop process is stable with the eigenvalues of* $A_\xi(\Lambda, K, L)$ *within the unit circle, and* $D_r(\Lambda, K, L) \nsubseteq D_r(I, K, L)$*, or (2) the attacked closed-loop process is such that* $\max_i |\lambda_i(A_\xi(\Lambda, K, L)| > 1$*, then the attack is potentially detectable with respect to the detection scheme in Eq. 3.10.*

*Proof.* The proof is divided into two parts. Part 1 considers the case when the attacked closed-loop process is stable, but $D_r(\Lambda, K, L) \nsubseteq D_r(I, K, L)$. Part 2 considers the case when the attack renders the closed-loop process unstable in the conventional sense such that $\max_i |\lambda_i(A_\xi(\Lambda, K, L))| > 1$.

Part 1: Consider that the attacked closed-loop process remains stable with the eigenvalues of $A_\xi(\Lambda, K, L)$ lying within the unit circle, and $D_r(\Lambda, K, L) \nsubseteq D_r(I, K, L)$. Since the origin is contained within the disturbance set (i.e., $0 \in F$), the origin is contained within the minimum invariant sets: $D_\xi(I, K, L)$ and $D_\xi(\Lambda, K, L)$ for the attack-free and attacked closed-loop process, respectively, from Eq. 3.7. If the disturbance is identically equal to 0 ($\mathbf{d} \equiv 0 \in \mathcal{F}$), the augmented state will be maintained at the origin for $\xi(0) = 0 \in D_\xi(\Lambda, K, L)$ implying that the residual of the attacked process is also maintained at the origin, which is within the attack-free terminal residual set, i.e., $r(t) = 0 \in D_r(I, K, L)$ for all $t \geq 0$. For such a realization of the disturbance and initial condition, the attack will go undetected for all $t \geq 0$. However, since $D_r(\Lambda, K, L) \not\subseteq D_r(I, K, L)$, $r(t) \in D_r(\Lambda, K, L) \setminus D_r(I, K, L)$ is possible for some $t \geq 0$, $\mathbf{d} \in \mathcal{F}$, and $\xi(0) \in D_r(\Lambda, K, L)$ following similar arguments as that used in the proof of Proposition 3. This implies that the attack is potentially detectable.

Part 2: Consider that the attacked closed-loop process is such that $\max_i|\lambda_i(A_\xi(\Lambda, K, L))| > 1$. Similar logic as that used in Part 1 may be applied to show that the attack is potentially detectable. If $\xi(0) = 0$ and $\mathbf{d} \equiv 0 \in \mathcal{F}$, the attack is not detected. On the other hand, since $D_\xi(\Lambda, K, L) = \mathbb{R}^{2n_x}$ by convention when the closed-loop process is rendered unstable by the attack, there exist $\xi(0) \in \mathbb{R}^{2n_x}$ such that $r(0) \notin D_r(I, K, L)$ and the attack is detected at $t = 0$. Therefore, the attack is potentially detectable.

□

**Proposition 15.** *Consider the closed-loop process with control system parameters $(K, L)$ under a multiplicative attack of magnitude $\Lambda \neq I$. Let the control system parameters $(K, L)$ stabilize the attack-free closed-loop process. If the attack renders the closed-loop process unstable in the sense that $\|\xi(t)\| \to \infty$ as $t \to \infty$ and the pair $(A_\xi(\Lambda, K, L), A_r(\Lambda))$ is observable, the attack is detected in finite time with respect to the detection scheme in Eq. 3.10.*

*Proof.* If the closed-loop process under attack is rendered unstable in the sense that $\|\xi(t)\| \to \infty$ as $t \to \infty$ and the pair $(A_\xi(\Lambda, K, L), A_r(\Lambda))$ is observable, the residuals are unbounded in the sense that $\|r(t)\| \to \infty$ as $t \to \infty$. This follows from Theorem 4 (B). Since the attack-free closed-loop process with control system parameters $(K, L)$ is stable, its minimum invariant set $D_\xi(I, K, L)$ is a compact (closed and bounded) set. As a result, the attack-free terminal residual set is also a compact set (from Eq. 3.9). There exists $R > 0$ such that $D_r(I, K, L) \subseteq B^{n_y}(R)$. Because the residuals of the attacked process are unbounded ($\|r(t)\| \to \infty$ as $t \to \infty$), for all $\epsilon > 0$ there exists $T > 0$ such that $\|r(t)\| > \epsilon$ for all $t > T$. Choosing $\epsilon > R$ shows there exists a finite time $T_1$ such that $\|r(T_1)\| > \epsilon > R$, which implies that $r(T_1) \notin D_r(I, K, L)$. Thus, the attack is detected in finite time and the attack is detectable. □

**Theorem 4.** *Consider the system*

$$\begin{aligned} z(t+1) &= A_z z(t) + B_\nu \nu(t) \\ \eta(t) &= C_z z(t) + D_\nu \nu(t) \end{aligned} \tag{B.1}$$

*with* $z(t) \in \mathbb{R}^{n_z}$*,* $\eta(t) \in \mathbb{R}^{n_\eta}$*, and* $\nu(t) \in \Gamma \subset \mathbb{R}^{n_\nu}$ *for all time* $t \geq 0$*, where* $\Gamma$ *is a compact set. If the pair* $(A_z, C_z)$ *is observable, and* $\|z(t)\| \to \infty$ *as* $t \to \infty$*, then* $\|\eta(t)\| \to \infty$ *as* $t \to \infty$*.*

*Proof.* Defining $\eta_n(t)$ and $\nu_n(t)$ as:

$$\eta_n(t) := \begin{bmatrix} \eta(t) \\ \vdots \\ \eta(t+n-1) \end{bmatrix}, \; \nu_n(t) := \begin{bmatrix} \nu(t) \\ \vdots \\ \nu(t+n-1) \end{bmatrix} \tag{B.2}$$

If the pair $(A_z, C_z)$ is observable, the observability matrix has rank $n_z$. Provided $\eta_n(t)$ and $\nu_n(t)$, $z(t)$ is the unique solution to the following system of equations if $(A_z, C_z)$ is observable:

$$\eta_n(t) = \underbrace{\begin{bmatrix} C_z \\ C_z A_z \\ \vdots \\ C_z A_z^{n-1} \end{bmatrix}}_{=:\mathcal{O}_n} z(t) + \underbrace{\begin{bmatrix} D_\nu & & & \\ C_z B_\nu & D_\nu & & \\ \vdots & \ddots & \ddots & \\ C_z A_z^{n-2} B_\nu & \cdots & C_z B_\nu & D_\nu \end{bmatrix}}_{=:\mathcal{B}_n} \nu_n(t) = \mathcal{O}_n z(t) + \mathcal{B}_n \nu_n(t) \tag{B.3}$$

where $\mathcal{O}_n$ is the observability matrix.

Since the pair $(A_z, C_z)$ is observable, $\mathcal{O}_n$ has full column rank and $\mathcal{O}_n^T \mathcal{O}_n$ is a positive definite matrix. Thus, $\|z\|_{\mathcal{O}_n^T \mathcal{O}_n} := \sqrt{z^T \mathcal{O}_n^T \mathcal{O}_n z}$ is a weighted Euclidean norm. Owing to the equivalence of norms, there exists $c > 0$ such that $\|z(t)\|_{\mathcal{O}_n^T \mathcal{O}_n} \geq c\|z(t)\|$. From Eq. B.3, the equivalence of norms, and the triangle inequality,

$$\begin{aligned} c\|z(t)\| &\leq \|\mathcal{O}_n z(t)\| = \|\eta_n(t) - \mathcal{B}_n \nu_n(t)\| \\ &\leq \|\eta(t)\| + \|\eta(t+1)\| + \ldots + \|\eta(t+n-1)\| + \|\mathcal{B}_n \nu_n(t)\| \end{aligned} \tag{B.4}$$

Since $\nu(t)$ is bounded for all $t \geq 0$, $\|\mathcal{B}_n \nu_n(t)\|$ is bounded. Because the last line of Eq. B.4 is a sum over a finite number of terms, $\|\eta(t)\| \to \infty$ as $t \to \infty$ if $\|z(t)\| \to \infty$ as $t \to \infty$. □

# Appendix C

# Proofs for Chapter 4

**Proposition 16.** *Consider the attack-free closed-loop process with* $(K, L)$*. If the matrix* $C$ *is invertible and* $\xi(0) \in D_\xi(I, K, L)$*, then the confidence region* $\Xi(K, L, t)$ *contains the augmented state, i.e.,* $\xi(t) \in \Xi(K, L, t)$*. Furthermore, the confidence region has a non-empty intersection with the minimum invariant set, i.e.,* $\Xi(K, L, t) \cap D_\xi(I, K, L) \neq \emptyset$*.*

*Proof.* This proposition is proved in two parts. In the first part, the containment of the augmented state within the confidence region is considered. In the second part, the intersection of the confidence region with the attack-free minimum invariant set is considered.

*Part 1:* From Eq. 4.16 and Eq. 4.17,

$$\begin{aligned}\Xi(K, L, t) = \tilde{C}^{-1}\left(\{\chi(t)\} \ominus \tilde{D}F\right) &= \left(\{\tilde{C}^{-1}\tilde{C}\xi(t)\} \oplus \{\tilde{C}^{-1}\tilde{D}d(t)\} \ominus \tilde{C}^{-1}\tilde{D}F\right) \\ &= \{\xi(t)\} \oplus \{\tilde{C}^{-1}\tilde{D}d(t)\} \ominus \tilde{C}^{-1}\tilde{D}F\end{aligned} \tag{C.1}$$

for the attack-free process. Because the process disturbances and measurement noise are bounded within the compact set $(F)$ containing the origin, the origin is contained in the set $\{\tilde{C}^{-1}\tilde{D}d(t)\} \ominus \tilde{C}^{-1}\tilde{D}F$. Therefore, the right-hand side of Eq. C.1 contains the augmented state of the attack-free process, and the confidence region constructed at any time $t \geq 0$ contains the augmented state, i.e., $\xi(t) \in \Xi(K, L, t)$.

*Part 2:* If the augmented state of the attack-free process at time $t = 0$ is contained within its minimum invariant set, then due to the forward invariance of the minimum invariant set, the augmented state is contained within the set for all time, i.e., $\xi(t) \in D_\xi(I, K, L)$

for all $t \geq 0$. From the proof of Part 1, the confidence region constructed for the attack-free process at any time contains the augmented state ($\xi(t) \in D_{\xi}(I, K, L)$). Therefore, $\Xi(K, L, t)$ and $D_{\xi}(I, K, L)$ both contain the augmented state $\xi(t)$, and have a non-empty intersection, i.e., $\Xi(K, L, t) \cap D_{\xi}(I, K, L) \neq \emptyset$.

□

**Proposition 17.** *Consider the closed-loop process with* $(K, L)$*. Let the matrix* $C$ *be invertible and* $\xi(0) \in D_{\xi}(I, K, L)$*. If the confidence region does not intersect with the minimum invariant set of the attack-free closed-loop process, i.e.,* $\Xi(K, L, t) \cap D_{\xi}(I, K, L) = \emptyset$*, then the process is not attack-free.*

*Proof.* This proposition is proved by contradiction. Assume that the closed-loop process is attack-free. From the proof of Part 1 of Proposition 6, $\xi(t) \in \Xi(K, L, t)$ at any time $t \geq 0$. If the confidence region does not intersect with the minimum invariant set of the process, i.e., $\Xi(K, L, t) \cap D_{\xi}(I, K, L) = \emptyset$, the minimum invariant set cannot contain the augmented state of the process, i.e., $\xi(t) \notin D_{\xi}(I, K, L)$. This is a contradiction, since, for the attack-free process, the augmented state is always contained within its minimum invariant set, i.e., $\xi(t) \in D_{\xi}(I, K, L)$ if $\xi(0) \in D_{\xi}(I, K, L)$. Thus, the process cannot be attack-free. □

**Theorem 5.** *Consider the closed-loop process with* $(K_1, L_1)$*. Let the matrix* $C$ *be invertible and* $\xi(0) \in D_{\xi}(I, K_1, L_1)$*. Assume that a controller-observer parameter switch from* $(K_1, L_1)$ *to* $(K_2, L_2)$ *occurs at* $t_s$*. If the closed-loop process is attack-free and the confidence region satisfies* $\Xi(K_1, L_1, t_s) \cap D_{\xi}(I, K_1, L_1) \subseteq D_{\xi}(I, K_2, L_2)$*, then no alarms are generated by the detection scheme of the form in Eq. 4.15. Furthermore, if there is an alarm generated by the detection scheme at some time* $t_d$*, then the closed-loop process is not attack-free.*

*Proof.* The proof is divided into two parts. In the first part, the attack-free process is considered. In the second part, the generation of an alarm is considered.

*Part 1:* Because $D_{\xi}(I, K_1, L_1)$ is a forward invariant set for the attack-free closed-loop process with $(K_1, L_1)$, for $t \in [0, t_s]$, the augmented state of the attack-free process is contained within $D_\xi(I, K_1, L_1)$. From Proposition 6, the augmented state of the attack-free process is contained within the intersection of the confidence region and the minimum invariant set, i.e., $\xi(t) \in \Xi(K_1, L_1, t) \cap D_\xi(I, K_1, L_1)$ for $t \in [0, t_s]$ when the matrix $C$ is invertible. If the intersection of the confidence region at $t_s$ and the minimum invariant set with $(K_1, L_1)$ is a subset or equal to the minimum invariant set of the attack-free process with $(K_2, L_2)$, i.e., $\Xi(K_1, L_1, t_s) \cap D_\xi(I, K_1, L_1) \subseteq D_\xi(I, K_2, L_2)$, the augmented state at $t_s$ is contained within the minimum invariant set of the attack-free process with $(K_2, L_2)$, i.e., $\xi(t_s) \in D_\xi(I, K_2, L_2)$. For this case, $\xi(t) \in D_\xi(I, K_2, L_2)$ for $t \geq t_s$ owing to the invariance of $D_\xi(I, K_2, L_2)$.

The value of the monitoring variable $\chi(t)$ will be within the corresponding terminal set for all $t \geq 0$. In particular, $\chi(t) \in D_\chi(I, K_1, L_1)$ for $t \in [0, t_s]$ and $\chi(t) \in D_\chi(I, K_2, L_2)$ for $t \geq t_s$ by construction of the sets $D_\chi(I, K_1, L_1)$ and $D_\chi(I, K_2, L_2)$. Hence, no alarms are generated with the detection scheme in Eq. 4.15 for the attack-free process if

$$\Xi(K_1, L_1, t_s) \cap D_\xi(I, K_1, L_1) \subseteq D_\xi(I, K_2, L_2) \tag{C.2}$$

*Part 2:* Consider the interval $[0, t_s]$ and let $\xi(0) \in D_\xi(I, K_1, L_1)$. If an alarm is generated for any $t_d \in [0, t_s]$, the value of the monitoring variable is not within its terminal set, i.e., $\chi(t_d) \notin D_\chi(I, K_1, L_1)$. By construction of $D_\chi(I, K_1, L_1)$, the closed-loop process is not attack-free. The attack is detected at $t_d$.

The remaining part is proved by contradiction. Specifically, consider the case that no alarms are raised for all $t \in [0, t_s]$. Let a parameter switch from $(K_1, L_1)$ to $(K_2, L_2)$ occur at $t_s \geq 0$ when the confidence region satisfies $\Xi(K_1, L_1, t_s) \cap D_\xi(I, K_1, L_1) \subseteq D_\xi(I, K_2, L_2)$. Assume that the process is attack-free for all $t \geq 0$. Let an alarm be generated at some time $t_d \geq t_s$, implying that the value of the monitoring variable at the time $t_d$ is not in the terminal set of the attack-free closed-loop process with $(K_2, L_2)$, i.e., $\chi(t_d) \notin D_\chi(I, K_2, L_2)$. When $\Xi(K_1, L_1, t_s) \cap D_\xi(I, K_1, L_1) \subseteq D_\xi(I, K_2, L_2)$, the process is attack-free, and $\xi(0) \in D_\xi(I, K_1, L_1)$, the monitoring variable evolves according to $\chi(t) \in D_\chi(I, K_1, L_1)$ for $t \in [0, t_s]$ and $\chi(t) \in D_\chi(I, K_2, L_2)$ for $t \geq 0$ by Part 1. Hence, no alarms can be generated. This leads to a contradiction. The closed-loop process is not attack-free when an attack is detected at any $t_d \geq 0$, $\xi(0) \in D_\xi(I, K_1, L_1)$, and

$\Xi(K_1, L_1, t_s) \cap D_\xi(I, K_1, L_1) \subseteq D_\xi(I, K_2, L_2).$ □

# Appendix D

# Proofs for Chapter 5

**Proposition 18.** *Consider the closed-loop process in Eq. 5.9 monitored by the reachable set-based detection scheme in Eq. 5.14, with an initial set $\mathcal{R}_0^\xi$. The closed-loop process is attack-free only if the output of the detection scheme in Eq. 5.14 is $h(\eta_k, \mathcal{R}_k^\xi) = 0$ for all $k \in \mathbb{Z}^+$.*

*Proof.* For the attack-free closed-loop process with $\xi_0 \in \mathcal{R}_0^\xi$, the augmented state is contained within the $k$-step reachable set, i.e., $\xi_k \in \mathcal{R}_k^\xi(\mathcal{R}_0^\xi)$ for all $k \in \mathbb{Z}^+$. From Eq. 5.13, the generalized monitoring variable of the attack-free process is contained within its $k$-step reachable set, i.e., $\eta_k \in \mathcal{R}_k^\eta(\mathcal{R}_k^\xi)$ for all $k \in \mathbb{Z}^+$. From Eq. 5.14, the output of the detection scheme is $h(\eta_k, \mathcal{R}_k^\xi) = 0$ for all $k \in \mathbb{Z}^+$. □

**Proposition 19.** *Consider the closed-loop process in Eq. 5.9, with an initial set $\mathcal{R}_0^\xi$, under an FDIA beginning at $k = 0$. The attack is undetectable with respect to the detection scheme in Eq. 5.14 and the initial set $\mathcal{R}_0^\xi$ if and only if the reachable sets of the monitoring variable for the attacked and the attack-free process satisfy $\mathcal{R}_k^{\eta_a}(\mathcal{R}_k^{\xi_a}) \subseteq \mathcal{R}_k^\eta(\mathcal{R}_k^\xi)$ for all $k \in \mathbb{Z}^+$.*

*Proof. Sufficiency*: Consider the attacked closed-loop process and the initial set $\mathcal{R}_0^\xi$. Let the reachable set of the monitoring variable for the attacked process be a subset of, or equal to, the reachable set for the attack-free process; i.e., $\mathcal{R}_k^{\eta_a}(\mathcal{R}_k^{\xi_a}) \subseteq \mathcal{R}_k^\eta(\mathcal{R}_k^\xi)$ for all $k \in \mathbb{Z}^+$. This implies that the monitoring variable values are contained within the reachable sets for the attack-free process ($\eta_k \in \mathcal{R}_k^{\eta_a}(\mathcal{R}_k^{\xi_a}) \subseteq \mathcal{R}_k^\eta(\mathcal{R}_k^\xi)$), and the detection scheme generates an output of $h(\eta_k, \mathcal{R}_k^{\xi}) = 0$ for all $k \in \mathbb{Z}^+$. Therefore, the attack is undetectable.

*Necessity*: Consider the attacked closed-loop process with the initial set $\mathcal{R}_0^{\xi}$. Let the FDIA begin at $k = 0$ and be undetectable with respect to the detection scheme in Eq. 5.14 and the initial set $\mathcal{R}_0^{\xi}$. By definition of an undetectable attack, $h(\eta_k, \mathcal{R}_k^{\xi}) = 0$ for all $k \in \mathbb{Z}^+$, $\xi_0 \in \mathcal{R}_0^{\xi}$, and $d_k \in \mathcal{D}$. From Eq. 5.14, this implies that $\eta_k \in \mathcal{R}_k^{\eta}(\mathcal{R}_k^{\xi})$ for all $k \in \mathbb{Z}^+$. However, the process is subjected to the FDIA, so $\eta_k \in \mathcal{R}_k^{\eta_a}(\mathcal{R}_k^{\xi_a})$ for all $k \in \mathbb{Z}^+$, implying that $\mathcal{R}_k^{\eta_a}(\mathcal{R}_k^{\xi_a}) \subseteq \mathcal{R}_k^{\eta}(\mathcal{R}_k^{\xi})$ for all $k \in \mathbb{Z}^+$. □

**Proposition 20.** *Consider the closed-loop process in Eq. 5.9, with an initial set $\mathcal{R}_0^{\xi}$, under an FDIA beginning at $k = 0$. The attack is detectable if the reachable sets of the monitoring variable for the attacked and the attack-free process satisfy $\mathcal{R}_k^{\eta_a}(\mathcal{R}_k^{\xi_a}) \cap \mathcal{R}_k^{\eta}(\mathcal{R}_k^{\xi}) = \emptyset$ for some $k \in \mathbb{Z}^+$.*

*Proof.* If the reachable sets of the generalized monitoring variable for the attacked and the attack-free process do not intersect at some $k \in \mathbb{Z}^+$, i.e., $\mathcal{R}_k^{\eta_a}(\mathcal{R}_k^{\xi_a}) \cap \mathcal{R}_k^{\eta}(\mathcal{R}_k^{\xi}) = \emptyset$, no value of the monitoring variable that is contained within the attacked reachable set is contained within the attack-free reachable set, i.e., $\eta_k \notin \mathcal{R}_k^{\eta}(\mathcal{R}_k^{\xi})$. The output of the detection scheme in this case is $h(\eta_k, \mathcal{R}_k^{\xi}) = 1$, and the attack is detected. Hence, the attack is detectable. □

# Appendix E

# Proofs for Chapter 6

**Proposition 21.** *Consider the attacked closed-loop process with $(K, L)$ and an initial set of states $\mathcal{R}_0^\xi$. Let an attack destabilize the process in the sense $\|\xi_t\| \to \infty$ as $t \to \infty$. If the matrix pair $(A^{\xi_a}(K, L), C^{\eta_a})$ is observable, then the attack is detectable with respect to the reachable set-based detection scheme in Eq. 6.15.*

*Proof.* For an attack-free closed-loop process with $(K, L)$ and a set of initial states $\mathcal{R}_0^\xi$, its $t$-step reachable set is a compact set. This follows from Eq. 6.16b, and the assumptions that the disturbances are bounded within a compact set ($\mathcal{D}$) and the set of initial states ($\mathcal{R}_0^\xi$) is a compact set. Let the attack cause the closed-loop process to be unstable in the sense that $\|\xi_t\| \to \infty$ as $t \to \infty$, i.e., there exists a $R > 0$ such that $\|\xi_t\| > R$ after some finite $t' > 0$. From Proposition 23, if the matrix pair $(A^{\xi_a}(K, L), C^{\eta_a})$ is observable, and if $\|\xi_t\| > R$, then at the time step $t' > 0$, the monitoring variable of the process cannot be bounded within any compact set. Meaning that at the time step $t'$, the monitoring variable is not bounded within the $t'$-step reachable set of the attack-free process ($\eta_{t'} \notin \mathcal{R}_{t'}^\eta(\mathcal{R}_{t'}^\xi(K, L))$), and the detection scheme in Eq. 6.12 detects the attack at the time step $t' > 0$ with an output of 1. Therefore, the attack is detectable with respect to the reachable set-based detection scheme in Eq. 6.15. □

**Proposition 22.** *Consider the attack-free closed-loop process under the nominal mode with an initial set of states $\mathcal{R}_0^\xi$, which is monitored by the reachable set-based detection scheme in Eq. 6.18. Let multiple control mode switches between the nominal and the* *attack-sensitive mode be implemented on the process. Let the switching instances be randomly chosen time steps $t_{s_i} \in \mathbb{Z}^+$, where $i \in \{1, 2, 3, \ldots\}$ such that $t_{s_{i+1}} > t_{s_i}$. The reachable set-based detection scheme generates no alarms for all $t \in \mathbb{Z}^+$.*

*Proof.* Consider the attack-free closed-loop process operated under the nominal control mode with the set of initial states $\mathcal{R}_0^\xi$. Before a switch to the attack-sensitive control mode occurs, the reachable set-based detection scheme in Eq. 6.18 accounts for all values of the attack-free process and generates no false alarms. Consider the switch to the attack-sensitive control mode at the randomly chosen time step $t_{s_1} \in \mathbb{Z}^+$, followed by another switch back to the nominal control mode at the randomly chosen time step $t_{s_2} > t_{s_1}$. From Eq. 6.18, the reachable set-based detection scheme switches to monitoring the process based on the attack-free reachable sets computed for the attack-free process operated under the attack-sensitive mode, i.e., $(K_t, L_t) = (K^f, L^f)$ for all $t \in [t_{s_1}, t_{s_2})$. For the attack-free process operated under the attack-sensitive control mode, its reachable set at each time step after the switch contains all values of the monitoring variable ($\eta_t \in \mathcal{R}_t^\eta(\mathcal{R}_t^\xi)$), leading to an output of 0 for $t \in [t_{s_1}, t_{s_2})$. Similarly, after switching back to the nominal control mode at time step $t_{s_2}$, the reachable set-based detection scheme generates no false alarms. Similarly, for all subsequent switches between the nominal and the attack-sensitive control modes implemented on the process at randomly chosen time steps, it can be demonstrated that the detection scheme in Eq. 6.18 generates no alarms. □

**Proposition 23.** *Consider the process:*

$$\begin{aligned} \psi_{k+1} &= A^\psi \psi_k + B^\omega \omega_k \\ \mu_k &= C^\psi \psi_k + D^\omega \omega_k \end{aligned} \tag{E.1}$$

*where $\psi_k \in \mathbb{R}^{n_\psi}$ is the state, $\omega_k \in \Omega \subset \mathbb{R}^{n_\omega}$ is a bounded input, $\Omega$ is a compact set, and $\mu_k \in \mathbb{R}^{n_\mu}$ is the output. Let the pair $(A^\psi, C^\psi)$ be observable. For any compact set $\Gamma$, there exists $R > 0$ such that $\mu_i \notin \Gamma$ for some $i \in \{k, \ldots, k + n_\psi - 1\}$ if $\|\psi_k\| > R$.*

*Proof.* Defining $\mu_{k_{n_\psi}}$ and $\omega_{k_{n_\psi}}$ as the vectors consisting of states and input values over the last $n$ time steps:

$$\mu_{k_{n_\psi}} := \begin{bmatrix} \mu_k \\ \vdots \\ \mu_{k+n_\psi-1} \end{bmatrix}, \ \omega_{k_{n_\psi}} := \begin{bmatrix} \omega_k \\ \vdots \\ \omega_{k+n_\psi-1} \end{bmatrix} \tag{E.2}$$

If the pair $(A^\psi, C^\psi)$ is observable, $\psi_k$ is a unique solution to the system of equations:

$$\mu_{k_{n_\psi}} = \underbrace{\begin{bmatrix} C^\psi \\ C^\psi A^\psi \\ \vdots \\ C^\psi (A^\psi)^{n_\psi-1} \end{bmatrix}}_{=:\mathcal{O}_{n_\psi}} \psi_k + \underbrace{\begin{bmatrix} D^\omega & & \\ C^\psi B^\omega & D^\omega & \\ \vdots & \ddots & \ddots \\ C^\psi (A^\psi)^{n_\psi-2} B^\omega & \cdots \quad C^\psi B^\omega & D^\omega \end{bmatrix}}_{=:\mathcal{B}_{n_\psi}} \omega_{k_{n_\psi}} = \mathcal{O}_{n_\psi}\psi_k + \mathcal{B}_{n_\psi}\omega_{k_{n_\psi}} \tag{E.3}$$

where $\mathcal{O}_{n_\psi}$ is the observability matrix.

From Eq. E.3, $\|\mathcal{O}_{n_\psi}\psi_k\| = \|\mu_{k_{n_\psi}} - \mathcal{B}_{n_\psi}\omega_{k_{n_\psi}}\|$. Let $\|\psi\|_{\mathcal{O}^T_{n_\psi}\mathcal{O}_{n_\psi}} := \sqrt{\psi^T \mathcal{O}^T_{n_\psi}\mathcal{O}_{n_\psi}\psi} = \|\mathcal{O}_{n_\psi}\phi_k\|$. Note that $\|\cdot\|_{\mathcal{O}^T_{n_\psi}\mathcal{O}_{n_\psi}}$ is a norm because $(A^\psi, C^\psi)$ is observable, so $\mathcal{O}_{n_\psi}$ has full column rank and $\mathcal{O}^T_{n_\psi}\mathcal{O}_{n_\psi}$ is positive definite. From the equivalence of norms, there exists $c > 0$ such that $\|\psi_k\|_{\mathcal{O}^T_{n_\psi}\mathcal{O}_{n_\psi}} \geq c\|\psi_k\|$. From the triangle inequality,

$$c\|\psi_k\| \leq \|\psi_k\|_{\mathcal{O}^T_{n_\psi}\mathcal{O}_{n_\psi}} = \|\mu_{k_{n_\psi}} - \mathcal{B}_{n_\psi}\omega_{k_{n_\psi}}\| \leq \|\mu_{k+n_\psi-1}\| + \cdots + \|\mu_k\| + \|\mathcal{B}_{n_\psi}\omega_{k_{n_\psi}}\| \tag{E.4}$$

Since $\omega_k$ is bounded within a compact set $\Omega$, there exists $b > 0$ such that $\|\mathcal{B}_n\omega_{k_n}\| \leq b$ for all $\omega_{k_n} \in \Omega \times \cdots \times \Omega$. Using this bound and from Eq. E.4,

$$c\|\psi_k\| \leq \|\mu_{k+n_\psi-1}\| + \cdots + \|\mu_k\| + b \tag{E.5}$$

For any compact set $\Gamma$, there exists $r > 0$ such that $\|\mu\| > r$ implies that $\mu \notin \Gamma$. Let $\mathcal{J}$ be the set of indices of the terms in the right-hand side of Eq. E.5 that are strictly greater than $r$ ($\mathcal{J} \subseteq \{k, k+1, \ldots, k+n_\psi-1\}$). From Eq. E.5,

$$c\|\psi_k\| \leq (n_\psi - m)r + \sum_{j\in\mathcal{J}} \|\mu_j\| + b \tag{E.6}$$

where $m$ is the cardinality of $\mathcal{J}$. If $\|\phi_k\| > \frac{n_\psi r+b}{c}$ and from Eq. E.6,

$$mr < \sum_{j \in \mathcal{J}} \|\mu_j\| \tag{E.7}$$

Consider two cases: (1) $m = 0$, i.e., $\mathcal{J}$ is an empty set, and (2) $m \neq 0$. In the first case, Eq. E.7 leads to a contradiction when $m = 0$. Therefore, $m \neq 0$, implying $\|\mu_j\| > r$ for some $j \in \{k, k+1, \ldots, k+n_\psi - 1\}$ and $\mu_j \notin \Gamma$. Choosing $R > \frac{n_\psi r+b}{c}$ completes the proof.

□

# References

1. A. Alanqar, M. Ellis, and P. Christofides, “Economic model predictive control of nonlinear process systems using empirical models,” *AIChE Journal*, vol. 61, no. 3, pp. 816–830, 2015.

2. Critical infrastructure sectors. https://www.cisa.gov/topics/critical-infrastructure-security-and-resilience/critical-infrastructure-sectors. Date accessed: 02/10/2024.

3. Chemical sector. https://www.cisa.gov/topics/critical-infrastructure-security-and-resilience/critical-infrastructure-sectors/chemical-sector. Date accessed: 02/10/2024.

4. K. E. Hemsley and R. E. Fisher, “History of industrial control system cyber incidents,” Idaho National Lab., Idaho Falls, ID (United States), Tech. Rep. INL/CON-18-44111, 2018.

5. Official alerts & statements. https://www.cisa.gov/stopransomware/official-alerts-statements-cisa.

6. P. Vähäkainu, M. Lehto, and A. Kariluoto, *Cyberattacks Against Critical Infrastructure Facilities and Corresponding Countermeasures.* Cham: Springer International Publishing, 2022, pp. 255–292.

7. R. Setola, L. Faramondi, E. Salzano, and V. Cozzani, “An overview of cyber attack to industrial control system,” *Chemical Engineering Transactions*, vol. 77, pp. 907–912, 2019.

8. “Fact Sheet: Biden-harris administration expands public-private cybersecurity partnership to chemical sector,” https://www.whitehouse.gov/briefing-room/statements-releases/2022/10/26/fact-sheet-biden-harris-administration-expands-public-private-cybersecurity-partnership-to-chemical-sector/, Date accessed: 02/10/2024.

9. L. Tawalbeh, F. Muheidat, M. Tawalbeh, and M. Quwaider, “IoT privacy and security: Challenges and solutions,” *Applied Sciences*, vol. 10, no. 12, pp. 4102–4119, 2020.

10. H. Kayan, M. Nunes, O. Rana, P. Burnap, and C. Perera, “Cybersecurity of industrial cyber-physical systems: A review,” *ACM Computing Surveys*, vol. 54, no. 11s, pp. 1– 35, 2022.

11. I. M. Chapman, S. P. Leblanc, and A. Partington, “Taxonomy of cyber attacks and simulation of their effects,” in *Proceedings of the 2011 Military Modeling & Simulation Symposium.* Boston, Massachusetts: Society for Computer Simulation International, 3-7 April 2011, pp. 73 – 80.

12. B. Zhu, A. Joseph, and S. Sastry, “A taxonomy of cyber attacks on SCADA systems,” in *Proceedings of the 2011 International Conference on Internet of Things and 4th International Conference on Cyber, Physical and Social Computing*, Dalian, China, 19-22 October 2011, pp. 380–388.

13. T. H. Morris and W. Gao, “Industrial control system cyber attacks,” in *Proceedings of the 1st International Symposium for ICS & SCADA Cyber Security Research*, Leicester, UK, 16-17 September 2013, pp. 22–29.

14. A. E. Elhabashy, L. J. Wells, J. A. Camelio, and W. H. Woodall, “A cyber-physical attack taxonomy for production systems: A quality control perspective,” *Journal of Intelligent Manufacturing*, vol. 30, no. 6, pp. 2489–2504, 2019.

15. C. T. Lin, S. L. Wu, and M. L. Lee, “Cyber attack and defense on industry control systems,” in *Proceedings of the 2017 IEEE Conference on Dependable and Secure Computing*, Taipei, Taiwan, 7-10 August 2017, pp. 524–526.

16. S. Kim, G. Heo, E. Zio, J. Shin, and J. Song, “Cyber attack taxonomy for digital environment in nuclear power plants,” *Nuclear Engineering and Technology*, vol. 52, no. 5, pp. 995–1001, 2020.

17. Z. Drias, A. Serhrouchni, and O. Vogel, “Taxonomy of attacks on industrial control protocols,” in *Proceedings of the International Conference on Protocol Engineering (ICPE) and International Conference on New Technologies of Distributed Systems (NTDS)*, Paris, France, 22-24 July 2015, pp. 1–6.

18. H. T. Reda, A. Anwar, and M. Mahmood, “Comprehensive survey and taxonomies of false data injection attacks in smart grids: Attack models, targets, and impacts,” *Renewable and Sustainable Energy Reviews*, vol. 163, p. 112423, 2022.

19. G. Liang, S. R. Weller, J. Zhao, F. Luo, and Z. Y. Dong, “The 2015 Ukraine blackout: Implications for false data injection attacks,” *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 3317–3318, 2017.

20. M. Krotofil and A. A. Cárdenas, “Resilience of process control systems to cyber-physical attacks,” in *Proceedings of the 18th Nordic Conference on Secure IT Systems*, Ilulissat, Greenland, 18-21 October 2013, pp. 166–182.

21. C. Murguia and J. Ruths, “CUSUM and chi-squared attack detection of compromised sensors,” in *Proceedings of the 2016 IEEE Conference on Control Applications*, Buenos Aires, Argentina, 19-22 September 2016, pp. 474–480.

22. S. Narasimhan, N. H. El-Farra, and M. J. Ellis, “Detectability-based controller design screening for processes under multiplicative cyberattacks,” *AIChE Journal*, vol. 68, no. 1, p. e17430, 2022.

23. G. Na and Y. Eun, “A multiplicative coordinated stealthy attack and its detection for cyber physical systems,” in *Proceedings of the 2018 IEEE Conference on Control Technology and Applications*, Copenhagen, Denmark, 21-24 August 2018, pp. 1698–1703.

24. H. Durand, “State measurement spoofing prevention through model predictive control design,” in *Proceedings of the 6th IFAC Conference on Nonlinear Model Predictive Control*, vol. 51, Madison, WI, USA, 19-22 August 2018, pp. 543 – 548.

25. N. Hashemi and J. Ruths, “Codesign for resilience and performance,” *IEEE Transactions on Control of Network Systems*, vol. 10, no. 3, pp. 1387–1399, 2023.

26. S. Chen, Z. Wu, and P. D. Christofides, “A cyber-secure control-detector architecture for nonlinear processes,” *AIChE Journal*, vol. 66, no. 5, p. e16907, 2020.

27. H. Durand, “Anomaly-handling in Lyapunov-based economic model predictive control via empirical models,” in *Proceedings of the 21st IFAC World Congress*, vol. 53, Virtual, 11-17 July 2020, pp. 6911–6916.

28. F. Pasqualetti, F. Dörfler, and F. Bullo, “Attack detection and identification in cyber-physical systems,” *IEEE Transactions on Automatic Control*, vol. 58, no. 11, pp. 2715–2729, 2013.

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

30. L. Ye, F. Zhu, and J. Zhang, “Sensor attack detection and isolation based on sliding mode observer for cyber-physical systems,” *International Journal of Adaptive Control and Signal Processing*, vol. 34, no. 4, pp. 469–483, 2020.

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

32. S. Tan, J. M. Guerrero, P. Xie, R. Han, and J. C. Vasquez, “Brief survey on attack detection methods for cyber-physical systems,” *IEEE Systems Journal*, vol. 14, no. 4, pp. 5329–5339, 2020.

33. S. Chen, Z. Wu, and P. D. Christofides, “Cyber-attack detection and resilient operation of nonlinear processes under economic model predictive control,” *Computers & Chemical Engineering*, vol. 136, p. 106806, 2020.

34. B. Potteiger, Z. Zhang, and X. Koutsoukos, “Integrated instruction set randomization and control reconfiguration for securing cyber-physical systems,” in *Proceedings of the 5th Annual Symposium and Bootcamp on Hot Topics in the Science of Security*, Raleigh, North Carolina, 10-11 April 2018, pp. 1–10.

35. S. Z. Yong, M. Zhu, and E. Frazzoli, “Switching and data injection attacks on stochastic cyber-physical systems: Modeling, resilient estimation, and attack mitigation,” *ACM Transactions Cyber-Physical Systems*, vol. 2, no. 2, 2018.

36. T. Yucelen, W. M. Haddad, and E. Feron, “Adaptive control architectures for mitigating sensor attacks in cyber-physical systems,” *Cyber-Physical Systems*, vol. 2, no. 1-4, pp. 24–52, 2016.

37. A. A. Cárdenas, S. Amin, Z. S. Lin, Y. L. Huang, C. Y. Huang, and S. S. Sastry, “Attacks against process control systems: Risk assessment, detection, and response,” in *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*, Hong Kong, China, 22-24 March 2011, pp. 355–366.

38. C. Murguia and J. Ruths, “Characterization of a CUSUM model-based sensor attack detector,” in *Proceedings of the IEEE 55th Conference on Decision and Control*, Las Vegas, NV, USA, 12-14 December 2016, pp. 1303–1309.

39. ——, “On reachable sets of hidden CPS sensor attacks,” in *Proceedings of the 2018 American Control Conference*, Milwaukee, WI, USA, 27-29 June 2018, pp. 178–184.

40. H. Oyama and H. Durand, “Integrated cyberattack detection and resilient control strategies using Lyapunov-based economic model predictive control,” *AIChE Journal*, vol. 66, no. 12, p. e17084, 2020.

41. A. Tsuchiya, F. Fraile, I. Koshijima, A. Ortiz, and R. Poler, “Software defined networking firewall for industry 4.0 manufacturing systems,” *Journal of Industrial Engineering and Management*, vol. 11, no. 2, pp. 318–333, 2018.

42. H. Durand, “Process/equipment design implications for control system cybersecurity,” in *Proceedings of the 9th International Conference on Foundations of Computer-Aided Process Design*, vol. 47, Copper Mountain Resort Colorado, CO,USA, 14–18 July 2019 2019, pp. 263–268.

43. H. Wen, F. Khan, S. Ahmed, S. Imtiaz, and S. Pistikopoulos, “Risk assessment of human-automation conflict under cyberattacks in process systems,” *Computers & Chemical Engineering*, vol. 172, p. 108175, 2023.

44. S. Parker, Z. Wu, and P. D. Christofides, “Cybersecurity in process control, operations, and supply chain,” *Computers & Chemical Engineering*, vol. 171, p. 108169, 2023.

45. K. K. Rangan, H. C. Oyama, and H. Durand, “Integrated cyberattack detection and handling for nonlinear systems with evolving process dynamics under Lyapunov-based economic model predictive control,” *Chemical Engineering Research and Design*, vol. 170, pp. 147–179, 2021.

46. H. Oyama, K. K. Rangan, and H. Durand, “Handling of stealthy sensor and actuator cyberattacks on evolving nonlinear process systems,” *Journal of Advanced Manufacturing and Processing*, vol. 3, no. 3, p. e10099, 2021.

47. H. Oyama, D. Messina, K. K. Rangan, and H. Durand, “Lyapunov-based economic model predictive control for detecting and handling actuator and simultaneous sensor/actuator cyberattacks on process control systems,” *Frontiers in Chemical Engineering*, vol. 4, p. 810129, 2022.

48. S. Narasimhan, N. H. El-Farra, and M. J. Ellis, “A control-switching approach for cyberattack detection in process systems with minimal false alarms,” *AIChE Journal*, vol. 68, no. 12, p. e17875, 2022.

49. ——, “A reachable set-based scheme for the detection of false data injection cyberattacks on dynamic processes,” *Digital Chemical Engineering*, vol. 7, p. 100100, 2023.

50. ——, “Active multiplicative cyberattack detection utilizing controller switching for process systems,” *Journal of Process Control*, vol. 116, pp. 64–79, 2022.

51. A. Zedan and N. H. El-Farra, “A machine-learning approach for identification and mitigation of cyberattacks in networked process control systems,” *Chemical Engineering Research and Design*, vol. 176, pp. 102–115, 2021.

52. C. N. Mavridis, A. Kanellopoulos, K. G. Vamvoudakis, J. S. Baras, and K. H. Johansson, “Attack identification for cyber-physical security in dynamic games under cognitive hierarchy,” in *Proceedings of the 2023 IFAC World Congress*, Yokohama, Japan, 9-14 July 2023, pp. 11 223 – 11 228.

53. L. F. Cómbita, N. Quijano, and Á. A. Cárdenas, “On the stability of cyber-physical control systems with sensor multiplicative attacks,” *IEEE Access*, vol. 10, pp. 39 716–39 728, 2022.

54. H. Liu, Y. Mo, and K. H. Johansson, “Active detection against replay attack: A survey on watermark design for cyber-physical systems,” in *Lecture Notes in Control and Information Sciences*. Springer, 2021, pp. 145–171.

55. D. Zhang, Q. G. Wang, G. Feng, Y. Shi, and A. V. Vasilakos, “A survey on attack detection, estimation and control of industrial cyber–physical systems,” *ISA Transactions*, vol. 116, pp. 1–16, 2021.

56. K. Garg, R. G. Sanfelice, and Á. A. Cárdenas, “Control barrier function-based attack-recovery with provable guarantees,” in *Proceedings of the IEEE 61st Conference on Decision and Control*, Cancun, Mexico, 6 - 9 December 2022, pp. 4808 – 4813.

57. F. Akbarian, W. Tärneberg, E. Fitzgerald, and M. Kihl, “A security framework in digital twins for cloud-based industrial control systems: Intrusion detection and mitigation,” in *Proceedings of the 26th IEEE International Conference on Emerging Technologies and Factory Automation*, Västerås, Sweden, 7-10 Sep 2021, pp. 1–8.

58. Y. Hu, H. Li, H. Yang, Y. Sun, L. Sun, and Z. Wang, "Detecting stealthy attacks against industrial control systems based on residual skewness analysis," *European Association for Signal Processing Journal on Wireless Communications and Networking*, vol. 2019, no. 1, pp. 1–14, 2019.

59. H. R. Ghaeini, N. O. Tippenhauer, and J. Zhou, "Zero residual attacks on industrial control systems and stateful countermeasures," in *Proceedings of the 14th International Conference on Availability, Reliability and Security*, Canterbury, CA, United Kingdom, 26-29 August 2019, pp. 1–10.

60. S. Chen, Z. Wu, and P. D. Christofides, "Cyber-attack detection and resilient operation of nonlinear processes under economic model predictive control," *Computers & Chemical Engineering*, vol. 136, p. 106806, 2020.

61. Z. Wu and P. D. Christofides, *Process Operational Safety and Cybersecurity: A Feedback Control Approach.* Springer, 2021.

62. S. Weerakkody, O. Ozel, P. Griffioen, and B. Sinopoli, "Active detection for exposing intelligent attacks in control systems," in *Proceedings of the IEEE Conference on Control Technology and Applications*, Hawai'i, USA, 27-30 Aug 2017, pp. 1306–1312.

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

64. T. Huang, B. Satchidanandan, P. R. Kumar, and L. Xie, "An online detection framework for cyber attacks on automatic generation control," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 6816–6827, 2018.

65. Y. Hu, H. Li, H. Yang, Y. Sun, L. Sun, and Z. Wang, "Detecting stealthy attacks against industrial control systems based on residual skewness analysis," *European Association for Signal Processing Journal on Wireless Communications and Networking*, vol. 2019, no. 1, pp. 1–14, 2019.

66. Z. Chu, J. Zhang, O. Kosut, and L. Sankar, “Unobservable false data injection attacks against PMUs: Feasible conditions and multiplicative attacks,” in *Proceedings of the IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids*, Aalborg, Denmark, 29-31 October 2018, pp. 1–6.

67. A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, “Revealing stealthy attacks in control systems,” in *Proceedings of the 50th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, Illinois, USA, 1-5 October 2012, pp. 1806–1813.

68. S. Narasimhan, N. H. El-Farra, and M. J. Ellis, “Controller switching-enabled active detection of multiplicative cyberattacks on process control systems,” in *Proceedings of the American Control Conference*, Atlanta, Georgia, 8-10 June 2022, pp. 2473–2478.

69. C. Trapiello and V. Puig, “Input design for active detection of integrity attacks using set-based approach,” in *Proceedings of the IFAC World Congress*, Berlin, Germany, 12-17 July 2020, pp. 11 094–11 099.

70. S. Narasimhan, N. H. El-Farra, and M. J. Ellis, “A reachable set-based cyberattack detection scheme for dynamic processes,” in *Proceedings of the American Control Conference*, San Diego, CA, 31 May - 2 June 2023, pp. 3777–3782.

71. S. Narasimhan, M. J. Ellis, and N. H. El-Farra, “Detection of multiplicative false data injection cyberattacks on process control systems via randomized control mode switching,” *Processes*, vol. 12, no. 2, 2024.

72. S. Narasimhan, N. H. El-Farra, and M. J. Ellis, “A set-based control mode selection approach for active detection of false data injection cyberattacks,” in *Proceedings of the American Control Conference*, Toronto, Canada, 10-12 July 2024, p. In Press.

73. P. D. Christofides and N. H. El-Farra, *Control of Nonlinear and Hybrid Process Systems: Designs for Uncertainty, Constraints and Time-Delays.* Springer Science & Business Media, 2005, vol. 324.

74. D. Popescu, A. Gharbi, D. Stefanoiu, and P. Borne, *Process Control Design for Industrial Applications*. John Wiley & Sons, 2017.

75. J. Romagnoli and A. Palazoglu, *Introduction to Process Control*. CRC press, 2020.

76. S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

77. V. M. Kuntsevich and B. N. Pshenichnyi, "Minimal Invariant Sets of Dynamic Systems with Bounded Disturbances," *Cybernetics and Systems Analysis*, vol. 32, no. 1, pp. 58–64, 1996.

78. M. Mansouri, M. Sheriff, R. Baklouti, M. Nounou, H. Nounou, A. B. Hamida, and N. Karim, "Statistical fault detection of chemical process-comparative studies," *Journal of Chemical Engineering & Process Technology*, vol. 7, no. 1, pp. 282–291, 2016.

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

80. I. V. Kolmanovsky and E. G. Gilbert, "Theory and computation of disturbance invariant sets for discrete-time linear systems," *Mathematical Problems in Engineering*, vol. 4, pp. 317–367, 1998.

81. D. Q. Mayne and W. R. Schroeder, "Robust time-optimal control of constrained linear systems," *Automatica*, vol. 33, no. 12, pp. 2103–2118, 1997.

82. M. Herceg, M. Kvasnica, C. N. Jones, and M. Morari, "Multi-Parametric Toolbox 3.0," in *Proceedings of the European Control Conference*, Zürich, Switzerland, July 17–19 2013, pp. 502–510.

83. P. M. Frank and X. Ding, "Survey of robust residual generation and evaluation methods in observer-based fault detection systems," *Journal of Process Control*, vol. 7, no. 6, pp. 403–424, 1997.

84. S. Simani, C. Fantuzzi, and R. Patton, *Model-based fault diagnosis in dynamic systems using identification techniques.* Springer: London, 2003.

85. M. Blanke, M. Kinnaert, J. Lunze, and M. Staroswiecki, *Diagnosis and Fault-Tolerant Control.* Springer-Verlag Berlin Heidelberg, 2003.

86. R. Isermann, *Fault-Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance.* Springer-Verlag Berlin Heidelberg, 2006.

87. V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. N. Kavuri, "A review of process fault detection and diagnosis: Part I: Quantitative model-based methods," *Computers & Chemical Engineering*, vol. 27, pp. 293–311, 2003.

88. International Society of Automation, "ANSI/ISA-18.2-2016: Management of alarm systems for the process industries," International Society of Automation, Standard, 2009.

89. S. Narasimhan, N. H. El-Farra, and M. J. Ellis, "Active multiplicative cyberattack detection utilizing controller switching for process systems," *Journal of Process Control*, vol. 116, pp. 64–79, August 2022.

90. A. Girard, C. L. Guernic, and O. Maler, "Efficient computation of reachable sets of linear time-invariant systems with inputs," in *Proceedings of the International Workshop on Hybrid Systems: Computation and Control*, Santa Barbara, CA, 29-31 March 2006, pp. 257–271.

91. N. Hashemi, E. V. German, J. Pena Ramirez, and J. Ruths, "Filtering approaches for dealing with noise in anomaly detection," in *Proceedings of the IEEE 58th Conference on Decision and Control*, Nice, France, 11-13 December 2019, pp. 5356–5361.

92. V. Renganathan, B. J. Gravell, J. Ruths, and T. H. Summers, "Anomaly detection under multiplicative noise model uncertainty," *IEEE Control Systems Letters*, vol. 6, pp. 1873–1878, 2022.

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

94. W.-H. Ko, B. Satchidanandan, and P. R. Kumar, "Dynamic watermarking-based defense of transportation cyber-physical systems," *ACM Transactions on Cyber-Physical Systems*, vol. 4, no. 1, pp. 1–21, 2019.

95. H. Durand, "A nonlinear systems framework for cyberattack prevention for chemical process control systems," *Mathematics*, vol. 6, p. 169, 2018.

96. H. Oyama, D. Messina, K. K. Rangan, F. L. Akkarakaran, K. Nieman, H. Durand, K. Tyrrell, K. Hinzman, and M. Williamson, "Development of directed randomization for discussing a minimal security architecture," *Digital Chemical Engineering*, vol. 6, p. 100065, 2023.

97. L. J. Guibas, A. T. Nguyen, and L. Zhang, "Zonotopes as bounding volumes," in *ACM-SIAM Symposium on Discrete Algorithms*, vol. 3, Baltimore, Maryland, USA, 12 - 14 January 2003, pp. 803–812.

98. M. Althoff, G. Frehse, and A. Girard, "Set propagation techniques for reachability analysis," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 4, no. 1, pp. 369–395, 2021.

99. H. Lin and P. J. Antsaklis, "Stability and stabilizability of switched linear systems: a survey of recent results," *IEEE Transactions on Automatic Control*, vol. 54, no. 2, pp. 308–322, 2009.

100. M. Attar and W. Lucia, "Data-driven robust backward reachable sets for set-theoretic model predictive control," *IEEE Control Systems Letters*, vol. 7, pp. 2305–2310, 2023.

101. B. Savković, “Low complexity parameterized approximations of reachable sets for LTI systems,” in *Proceedings of the IEEE International Conference on Control and Automation*, Christchurch, New Zealand, 9-11 December 2009, pp. 960–965.

102. M. Althoff, “An introduction to CORA 2015,” in *Proceedings of the Workshop on Applied Verification for Continuous and Hybrid Systems*, vol. 34, Seattle, Washington, 13 April 2015, pp. 120–151.

103. S. Raković and K. I. Kouramas, “The minimal robust positively invariant set for linear discrete time systems: Approximation methods and control applications,” in *Proceedings of the IEEE 45th Conference on Decision and Control*, San Diego, CA, USA, 13-15 December 2006, pp. 4562–4567.

104. H. Oyama, D. Messina, K. K. Rangan, F. L. Akkarakaran, K. Nieman, H. Durand, K. Tyrrell, K. Hinzman, and W. Williamson, “Development of directed randomization for discussing a minimal security architecture,” *Digital Chemical Engineering*, vol. 6, p. 100065, 2023.

105. C. Trapiello and V. Puig, “Input design for active detection of integrity attacks using set-based approach,” in *Proceedings of the 21st IFAC World Congress*, Berlin, Germany, 11-17 July 2020, pp. 11 094–11 099.

106. M. Porter, P. Hespanhol, A. Aswani, M. Johnson-Roberson, and R. Vasudevan, “Detecting generalized replay attacks via time-varying dynamic watermarking,” *IEEE Transactions on Automatic Control*, vol. 66, no. 8, pp. 3502–3517, 2021.

107. P. Hespanhol, M. Porter, R. Vasudevan, and A. Aswani, “Dynamic watermarking for general LTI systems,” in *Proceedings of the IEEE 56th Conference on Decision and Control*, Melbourne, Australia, 12-15 December 2017, pp. 1834–1839.

108. A. Naha, A. Teixeira, A. Ahlén, and S. Dey, “Quickest physical watermarking-based detection of measurement replacement attacks in networked control systems,” *European Journal of Control*, vol. 71, p. 100804, 2023.

109. N. Babadi and A. Doustmohammadi, “A moving target defence approach for detecting deception attacks on cyber-physical systems,” *Computers and Electrical Engineering*, vol. 100, p. 107931, 2022.

110. Y. Hu, P. Xun, P. Zhu, Y. Xiong, Y. Zhu, W. Shi, and C. Hu, “Network-based multidimensional moving target defense against false data injection attack in power system,” *Computers & Security*, vol. 107, p. 102283, 2021.

111. M. Ghaderi, K. Gheitasi, and W. Lucia, “A blended active detection strategy for false data injection attacks in cyber-physical systems,” *IEEE Transactions on Control of Network Systems*, vol. 8, no. 1, pp. 168–176, 2021.

112. L. Montejano, “Some results about minkowski addition and difference,” *Mathematika*, vol. 43, pp. 265–273, 1996.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346 USA