The Pennsylvania State University
The Graduate School
College of Engineering

DISTRIBUTED CONTROL OF CYBER-PHYSICAL SYSTEMS: SECURITY, ECONOMICS AND SMART GRID

A Dissertation in
Electrical Engineering
by
Hunmin Kim

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2018

The dissertation of Hunmin Kim was reviewed and approved* by the following:

Minghui Zhu
Dorothy Quiggle Assistant Professor of The School of Electrical Engineering and Computer Science
Dissertation Advisor, Chair of Committee

Constantino Lagoa
Professor of The School of Electrical Engineering and Computer Science

Nilanjan Ray Chaudhuri
Assistant Professor of The School of Electrical Engineering and Computer Science

Asok Ray
Distinguished Professor of The Department of Mechanical and Nuclear Engineering

Kultegin Aydin
Department Head of Electrical Engineering

*Signatures are on file in the Graduate School.

Abstract

We have witnessed rapid emergence of cyber-physical systems (CPS), which integrate control systems with advanced technologies of sensing, computation and communication. Many CPS consist of a population of agents which operate in heterogeneous spatial and temporal scales, and interact with one another in various ways. Thus, it is mandatory to develop practical distributed control methodologies for agents, which provide autonomous decision making given local information, while guaranteeing satisfactory network-wide performance. This dissertation contributes to the broad field of distributed control of CPS and investigates three emerging problems: cyber-physical security, cyber-physical economics, and smart grid.

Cyber-physical security. CPS inherit the vulnerabilities of Information and Communications Technology (ICT) systems to cyber-attacks. Attackers can bypass existing cyber defenses and cause irreparable damage to the physical world. We study two specific problems: attack-resilient estimation and attack-resilient machine learning. In the attack-resilient estimation, we consider the scenario that switched nonlinear stochastic systems are threatened by both of signal attacks and switching attacks. The attack-resilient estimation problem is formulated as the simultaneous estimation of states, attack vectors and hidden modes. We propose a multi-mode algorithm where each mode is associated with an estimator and the most likely mode is chosen to generate the estimates of system states and attacker vectors. It is formally proven that the estimation errors of states and attack vectors of the true mode satisfy Practically Exponentially Stable in probability (PESp) like properties, when the hidden mode is fixed but remains unknown. Lastly, the developed theory is validated by numerical simulations on power systems and real-world experiments on mobile robots and connected vehicles.

In the attack-resilient machine learning, we aim to learn the system function of a nonlinear stochastic system subject to signal injection attacks. The problem is formulated as to estimate state and attack vectors as well as the unknown system function. In the proposed algorithm, Gaussian process regression (GPR) is utilized

to learn the system function using estimated states and attack vectors. The function estimates are then used to generate the estimates of state and attack vectors using input and state estimation technique. We show that the estimation errors of states and attack vectors satisfy PESp-like property, average case learning errors of the function approximation are diminishing if the number of state estimates whose estimation errors are non-zero is bounded. The developed algorithm is applied to power systems to demonstrate its performance.

Cyber-physical economics. In some CPS, agents act to maximize their own interests. It is interesting to bridge the gap between social welfare and individual interests. We propose a bi-level lottery, where a social planner at the high level announces a reward and, sequentially, agents at the low level jointly find a Nash equilibrium in response to the reward. We introduce user's heterogeneity parameter and social planner's perturbation parameter, and formulate an optimal bi-level lottery design problem where the Nash equilibrium of the lottery game is coincident with the socially optimal payoff or a greater payoff with least reward and perturbations. Through the analytical results on Nash equilibrium, we derive a convex approximation of the optimal bi-level lottery design problem, and identify that the approximation is exact under mild sufficient conditions. We verify the results via a case study on demand response in smart grid.

Smart grid. We investigate distributed frequency control of multi-machine power systems with unknown net loads. Proposed distributed frequency controllers leverage synchronous generators and demand response to handle fast changing and hard-to-predict net loads. In particular, local (adaptive) internal models reconstruct net loads, and reconstructed signals are assigned to synchronous generators and demand response to track or filter. We show that the system states are asymptotically convergent to the desired signals. Numerical simulations on the IEEE 68-bus test system as well as the Minni-WECC system demonstrate the effectiveness of the controllers and performance under a three-phase fault and load-switching during light/peak loads.

Table of Contents

List of Figures

List of Tables

Acknowledgments

This work would not have been possible without supports from Professor Minghui Zhu, who has been a great advisor. I am grateful to him for his patience and support in overcoming numerous obstacles faced in my research. He steered me in the right direction whenever I needed it. He also helped me to prepare for my future career and provided me valuable insights on research problems.

I also greatly appreciate to the committee members for their supports and productive discussion. Professor Constantino Lagoa provided valuable comments on overall problem formulation and theoretical analysis. Professor Nilanjan Ray Chaudhuri gave me insightful comments on power system analysis which is my major application. Professor Asok Ray discussed practical extensions in stochastic control. The discussion and comments would make my future research more solid and practical.

I would like to thank my principal collaborators. I would like to especially thank Professor Peng Liu for teaching me a lot about security and stimulating creative thinking. I also thanks Dr. Pinyao Guo for the productive discussions and for the help in intuitive presentations of theories. Thanks Dr. Jianming Lian who provided intuitions and valuable comments on the smart grid problem.

I am grateful to my colleagues in Professor Minghui Zhu's research group for the exciting discussions and supports. I would like to thank Mr. Yang Liu, Mr. Zhisheng Hu, and Mr. Guoxiang Zhao for the help and feedback.

Thanks my wife Jukoung Park and son Zion Kim for emotional supports throughout my life. Without their love and support, this work would not exist.

This dissertation was supported by the National Science Foundation (CNS-1505664), the Army Research Office (W911NF13-1-0421 and W911NF-15-1-0576), the Grid Modernization Initiative of the Department of Energy, and the College of Information Sciences and Technology at the Pennsylvania State University.

Dedication

I dedicate this dissertation to my beloved wife Jukoung and son Zion.

Chapter 1
Introduction

We have witnessed rapid emergence of cyber-physical systems (CPS), which integrate control systems with advanced technologies of sensing, computation and communication. CPS are uniquely featured by the strong coupling between the physical world and the cyber space. Such strong coupling enables highly distributed, complex and collaborative applications such as smart grid, self-driving cars, intelligent transportation systems, smart and connected communities and so on. The significance of CPS is emphasized in a report to President Obama [4]. Moreover, the US National Science Foundation (NSF) determines CPS as a key research area. We have experienced substantial progress in advancement of CPS technologies. This satisfies the continuously growing demands for specifications, and induces a driving force for a set of applications. There, however, is a lack of scientific and mathematical methodologies to control and optimize CPS, preventing efficient and effective management and realization of the projected applications. Thus, the opportunities for CPS are far-reaching and many challenges remain unsolved.

Distributed control is essential for many CPS. In particular, many CPS consist of a population of agents which operate in heterogeneous spatial and temporal scales, and interact with one another in various ways. Hence, it is mandatory to develop practical distributed control methodologies for agents, which provide autonomous decision-making given local information, while guaranteeing satisfactory network-wide performance. Distributed control has several advantages over the traditional centralized control methodology. First of all, it requires only local information obtained by communicating with neighboring agents. This feature enhances scalability to large-scale networked systems, and reduces the needs of com-

munications bandwidths. Secondly, decision-making and control are performed locally without any centralized coordination. As a consequence, computational burden is shared by multiple agents, and the systems are robust to failures of individual agents. The development of distributed control methodologies faces significant challenges as well. Firstly, CPS embed inherent complexities. In particular, CPS are composed of a huge number of agents who may have limited knowledge about global networks, although the agents' dynamics are coupled with others via physical and cyber layers. Even more, the agents may be heterogeneous and seek subobjectives which may not be aligned with network-wide goals. Secondly, CPS operate in dynamic, uncertain and even hostile environments. Environmental complexities could significantly degrade system performance and even cause mission failures. Distributed control of CPS has been receiving substantial attention [5–8]. Fundamental issues include consensus [9], distributed optimization [10, 11] and game-theoretic learning [12, 13]. Fundamental theory has been applied to power systems [14, 15], mobile robotic networks [16–18], sensor networks [19,20], smart buildings [21], etc. This dissertation contributes to this broad research area. In particular, we focus on three emerging problems: cyber-physical security, cyber-physical economics and smart grid.

1.1 Cyber-physical security

Cyber-physical systems (CPS) are emerging with the integration of traditional Information and Communications Technology (ICT) systems with physical components. Inherent vulnerabilities of ICT systems subject to cyber-attacks impose significant security risks on CPS. Therefore, it becomes a top-priority issue to protect CPS from cyber-attacks. In the security community, there have been abundant research on security of ICT systems. Based on the audit sources, traditional intrusion detection systems (IDS) can be categorized into two classes: host-based and network-based. Host-based IDS monitor [22, 23] in-and-out data packets and system files of a single host. Network-based IDS [24–26] check network traffic of a strategic point in the network. Since ICT systems are a part of CPS, traditional security techniques are necessary. However, they are not sufficient because they do not take into account physical systems of CPS. For example, traditional IDS for ICT systems monitor cyber-space misbehaviors only. Attacks launched through

physical channels do not trigger abnormal cyberspace behavior, and thus do not raise any alarm. Furthermore, traditional IDS usually require information of attack types and channels, which is difficult to obtain in advance, especially for zero-day attacks. New methodologies are needed to complement existing solutions for ICT systems. In CPS, regardless of the attack types and channels, the attacker targets physical systems and aims to abort the missions of physical systems or produce damage on them. In this case, the attacks can always be considered as external disturbances (or uncertainties) on physical systems. Control theory has a long history to deal with uncertainties such as robust control [27], adaptive control [28], stochastic control [29], etc. These control-theoretic methodologies provide valuable insights to ensure CPS security.

1.1.1 Attack-resilient estimation

Literature review. In CPS security, one research direction is to identify fundamental limitations of attack detectors. Paper [30] shows that sensor attacks against state estimation in electric power grids are able to induce arbitrary estimation errors, if injected signals are in the column space of the output matrix. It is shown in [31] that, for linear descriptor systems, signal injection attacks are not detectable if and only if attack signals excite zero dynamics. Another research direction is to design attack detectors against signal attacks and switching attacks. Papers [31, 32] formulate attack detection problems as ℓ_0/ℓ_∞ optimization problems. The problems, however, are non-convex and NP-hard in general [31]. To overcome the computational complexity, paper [32] proposes convex relaxations of the optimization problems. Paper [33] studies robustness of state estimator based on ℓ_0 optimization with respect to modeling errors (sampling, computation/actuation jitter, and synchronization). The Kalman filter is adopted in [34] to conduct attack-resilient state estimation in the presence of stochastic noises. A multi-modal Luenberger observer is designed in [35], where its memory usage increases linearly with the number of states and outputs. All aforementioned papers are limited to linear systems.

Contributions. In Chapter 2, we consider the scenario that switched nonlinear stochastic systems are threatened by both of signal attacks and switching attacks. The attack-resilient estimation problem is formulated as the simultaneous

estimation of states, attack vectors and hidden modes. We propose a multi-mode algorithm to solve the problem, where the algorithm associates an estimator to each mode and the estimators share the same structure. Each estimator calculates the estimates of states and attack vectors, recursively. The differences between obtained outputs and predicted outputs represent mode probabilities, and the mode estimator selects the most likely one. It is formally proven that the estimation errors of states and attack vectors of the true mode satisfy Practically Exponentially Stable in probability (PESp) like properties, when the hidden mode is fixed but remains unknown. Furthermore, we discuss a mode reduction method to reduce the computational complexity for switched linear stochastic systems, remaining the minimal number of modes to maintain the same detection capabilities as the power set. Towards our best knowledge, this is the first time that systematically studies unknown input, state, and mode estimation of switched nonlinear stochastic systems. In Chapter 3, we conduct numerical simulations on the IEEE 68-bus test system to demonstrate the effectiveness of the proposed algorithm on time-varying modes with a regular mode set and a reduced mode set. We also conduct real world experiments on mobile robots to show the performance of the proposed estimator. Lastly, we extend the results to distributed settings and conduct experiments on urban connected vehicles, where the vehicles collaboratively detect attacks.

1.1.2 Attack-resilient machine learning

Literature review. Attack-resilient machine learning is closely related to machine learning in the presence of training data errors. Different from our goal, related works (fault tolerant learning [37, 38], and adversarial learning [39, 40]) study probably approximately correct (PAC) learning in the presence of (malicious) noises. It has been shown in [37] that the random classification noise model is PAC-learnable. Robustness of PAC-learning algorithms is studied in [38]. The paper also shows that there is no tolerant algorithm in the presence of random attribute noises. Adversarial learning [39,40] studies worst-case error and robustness of learning algorithms when the attacker can choose a fixed random probability of errors. Since the papers focus on learning algorithms tolerating (malicious) errors rather than rejecting the effect of attacks, the performance of the machine learning

algorithms degrades in the presence of attacks.

Attack-resilient machine learning is also related to data-driven estimation of unknown dynamic systems using Gaussian process regression (GPR). GPR has been combined with extended Kalman filter [41], unscented Kalman filter [41, 42], and Cubature Kalman filter [43] to overcome a lack of knowledge of dynamic systems. However, no existing techniques can perform attack-resilient estimation or handle actuator attacks. No theoretic guarantee is provided in this set of papers.

Contributions. In Chapter 4, we consider to learn the system function of a partially unknown stochastic nonlinear system subject to signal injection attacks. The problem is formulated as a joint estimation of state, attack vectors, and system function estimation problem. We propose an algorithm which overcomes the limitation on knowledge of the system function by fusing Gaussian process regression with unknown input and state estimation technique. The system function is learned by GPR using estimated states and outputs. Then, the function estimates are incorporated to input and state estimation technique to obtain the estimates of state and attack vectors. It is formally proven that estimation errors of system states and attack vectors satisfy PESp-like property, and average case learning errors of system function approximation are diminishing if the number of state estimates whose estimation errors are non-zero is bounded. Numerical simulations on power systems show the effectiveness of the proposed algorithm.

The results of Chapters 2, 3 and 4 are based on the following publications.

(JP-1) **H. Kim**, P. Guo, M. Zhu, and P. Liu, Attack-resilient Estimation of Switched Nonlinear Stochastic Cyber-Physical Systems. In Automatica, Submitted

(CP-1) **H. Kim**, P. Guo, M. Zhu, and P. Liu, Attack-resilient Machine Learning of Dynamic Systems using Gaussian Process Regression. In the 2019 American Control Conference, In preparation.

(CP-2) P. Guo, **H. Kim**, N. Virani, J. Xu, M. Zhu, and P. Liu. RoboADS: Anomaly Detection against Sensor and Actuator Misbehaviors in Mobile Robots. In 2018 IEEE/IFIP International Conference on Dependable Systems and Networks, pages 574-585, 2018.

(CP-3) P. Guo, **H. Kim**, M. Zhu, and P. Liu. VCIDS: Collaborative Intrusion Detection of Sensor and Actuator Attacks on Connected Vehicles. In 13th

International Conference on Security and Privacy in Communication Networks, pages 377-396, 2017.

(CP-4) **H. Kim**, P. Guo, M. Zhu, and P. Liu, Attack-resilient Estimation of Switched Nonlinear Cyber-Physical Systems. In the 2017 American Control Conference, pages 4328-4333, 2017.

1.2 Cyber-physical economics

In many practical scenarios, control authorities in CPS are non-cooperative and seek for heterogeneous (or even conflicting) subobjectives. For example, power consumers compete over limited power generations, and Internet users share network bandwidths. This leads to competitions over limited resources and the degradation of network-wide performance. It is vital to eliminate the gaps between individual interests and network-wide goals such that the networked system can operate at an optimal level.

Literature review. To address the above issue, a common practice is to adopt methodologies in microeconomics and design mechanisms to align agents' preferences via side payments/pricing with social welfare. In an auction, bidders submit bids of the items, and an auctioneer sequentially determines item price and allocation. Vickrey-Clarke-Groves (VCG) auction [44–46] is the most well-known mechanism, and it has been shown that VCG is efficient and incentive compatible [47, 48]. Contract theory [49] is another type and has purpose of construction of a contract in the presence of asymmetric information of the agents. Contract theory has three types of models; i.e., moral hazard [50] (hidden information after the contract), adverse selection [51] (hidden information before the contract), and signaling [52] (some credential information provided by the agents). Optimal taxation [53] and trading [54] are other classes of mechanism design. Incentive design has two dynamic extensions. Algorithmic mechanism design is directed incentive design where principal incentivizes agents to follow specified algorithms or solve computation problems [55–57]. Incentive control is indirected where agents' choses are affected by rewards/prices determined by principal [58, 59].

Contributions. In some CPS, agents act to maximize their own interests. It is interesting to bridge the gap between social welfare and individual interests. In

Chapter 5, we propose a bi-level lottery, where a social planner at the high level announces a reward and, sequentially, agents at the low level jointly find a Nash equilibrium in response to the reward. It has been known that competitions among the agents results in efficiency losses in current lottery schemes, and social optimum is coincident with Nash equilibrium only when an infinite reward is given [60]. To mitigate the issue, we introduce user's heterogeneity parameter and social planner's perturbation parameter, and formulate an optimal bi-level lottery design problem where the Nash equilibrium of the lottery game is coincident with the socially optimal payoff or a greater payoff with least reward and perturbations. We formally analyze the properties of low-level Nash equilibrium, the price of anarchy, and the behavior of public goods and Nash equilibrium with respect to the reward and perturbations. Through the analytical results, we derive a convex approximation of the optimal bi-level lottery design problem, and identify that the approximation is exact under mild sufficient conditions. We verify the results via a case study on demand response in smart grid.

The results of Chapter 5 are based on the following publications.

(JP-2) **H. Kim** and M. Zhu. Optimal Bi-level Lottery Design for Multi-agent Networks. In Automatica, Submitted.

(CP-5) **H. Kim** and M. Zhu. Optimal Incentive Design for Distributed Stabilizing Control of Nonlinear Dynamic Networks. In IEEE Conference on Decision and Control, pages 2289-2294, 2015.

1.3 Smart grid

Smart grid is an important application of CPS. Centralized generating facilities are being distributed integrating small energy resources; e.g., photovoltaic systems, fuel cells, storage and electric vehicles. This tendency is likely to accelerate. Worldwide 144 countries now have their own political targets for increasing shares of renewable energy generations and especially European Union targets 20% of renewable energy shares by 2020 [61]. Integrating new technologies to power grid allows a more flexible and efficient management. However, renewable generation as well as other small energy resources are hard to predict, and an increasing proliferation imposes significant challenges to the operation and management of the

power grid. It becomes imperative to maintain grid stability and reliability despite the disturbances induced by renewable generation.

Literature review. There are many efforts made to control of the power grid under a variety of external disturbances. Representative techniques include Riccati equation [62], and H_2/H_∞ control [63, 64]. This set of papers aims to attenuate external disturbance; i.e., the impacts of external disturbances are reduced but not completely eliminated. On the other hand, disturbance rejection instead pursues to completely eliminate external disturbances and recover perfect stability. There are limited literature on disturbance rejection. Paper [65] develops distributed internal model controllers to ensure optimal frequency synchronization despite uncertain and time-varying loads.

Contributions. In Chapter 6, we study the frequency control of multi-machine power systems subject to uncertain and dynamic net loads, where the net loads contains power loads and renewable generation. Under the assumption that each net load consists of a set of sinusoidal functions, we design distributed controllers for the following two cases: (1) *robust adaptive frequency control*, where the frequencies of net loads are unknown; (2) *robust frequency control*, where the frequencies of net loads are known, with general dynamic systems. The proposed controllers reconstruct unknown disturbance signals using internal model, and utilize synchronous generators and demand response to filter them out. We formally show the stability of frequency via Lyapunov analysis and conduct numerical simulations on the IEEE 68-bus power system and Minni-WECC system. To our best knowledge, it is the first time to study distributed adaptive internal model control to handle external disturbances with unknown frequencies. As a byproduct of the analysis, we develop a new distributed constrained small-gain theorem, which is interesting on its own.

The results of Chapter 6 are based on the following publications.

(JP-3) **H. Kim**, M. Zhu, and J. Lian. Distributed Robust Adaptive Frequency Control of Power Systems with Dynamic Loads. In IEEE Transactions on Automatic Control, provisionally accepted.

(CP-6) **H. Kim** and M. Zhu. Distributed Robust Frequency Regulation of Smart Power Grid with Renewable Integration. In American Control Conference, pages 2347-2352, 2015.

Chapter 2
Cyber-physical security: Attack-resilient estimation (Theory)

2.1 Introduction

Cyber-Physical Systems (CPS) are systems which integrate control systems with advanced technologies of sensing, computation and communication. Security is of vital importance for CPS. Especially, because of the couplings between the cyber layer and physical layer, CPS bear vulnerabilities to cyberattacks which may cause irreparable damage to the physical layer [66]. For example, a natural gas flow control system in Russia was temporarily seized in 2000 and a sewage control system in Australia was attacked in the same year [67]. According to early studies on CPS security, the types of possible attacks on CPS can be categorized into signal attacks and switching attacks. Signal attacks include sensor attacks which tamper with sensor readings and actuator attacks which tamper with control commands. While signal attacks modify the magnitudes or timings [68, 69] of signals, switching attacks alter system structures [36, 70]. The attacks can be launched via communication jamming and malware; e.g., Trojan.

As discussed in Chapter 1, control-theoretic approaches to CPS security attract lots of attention. Papers [30, 33–35] handle sensor attacks only, and papers [31, 32] handle both sensor and actuator attacks. Recent paper [36] designs an attack-resilient estimator for stochastic linear systems when there are sensor attacks, actuator attacks, and switching attacks. All aforementioned papers are limited to linear systems.

Our attack-resilient estimator design method is based on simultaneous unknown Input and State Estimation (ISE). Early research of this area focuses on state estimation without estimating unknown inputs [71, 72]. Unbiased and minimum variance unknown input and state estimators are designed for linear systems without direct feedthrough matrix [73] and with full-column rank direct feedthrough matrix [74, 75], and with rank-deficient direct feedthrough matrix [76]. Noticeably, this set of papers is restricted to linear systems.

Chapter organization. A motivating example of CPS model and attack model is introduced in Section 2.2. Section 2.3 introduces system model, attack model and defender's knowledge. Moreover, the state, attack vector, and mode estimation problem is formulated in the same section. We propose Nonlinear unknown Input, State and Mode Estimator (NISME) to solve the estimation problem in Section 2.4. The stability of the proposed estimator is formally analyzed in Section 2.5. Section 2.6 discusses justifications of the assumptions, connections to existing works, and a way to reduce computational complexity caused by unknown signal attack locations, for switched stochastic linear systems.

Notations. Given a vector a_k, we use $\hat{a}_k$ and $\tilde{a}_k$ to denote an estimate of a_k and induced estimation error $\tilde{a}_k = a_k - \hat{a}_k$, respectively. Its error covariance is defined by $P_k^a \triangleq \mathbb{E}[\tilde{a}_k \tilde{a}_k^T]$, and cross error covariance with b_k is $P_k^{ab} \triangleq \mathbb{E}[\tilde{a}_k \tilde{b}_k^T] = (P_k^{ba})^T$.

We use the following definition for filter stability of nonlinear systems.

Definition 2.1.1 *Stochastic process $x(t)$ is said to be Practically Exponentially Stable in probability (PESp) if for any $\gamma \in (0, 1)$, there exist positive constants α, b, c, and δ such that, for any $\|x(0)\| \leq \delta$, the following holds for all $t \geq 0$:*

$$P(\|x(t)\| < \alpha e^{-bt} \|x(0)\| + c) \geq 1 - \gamma.$$

PESp is a special case of stochastic input-to-state stability [77] when input is absent and class $\mathcal{KL}$ function is exponential in t and linear in $\|x(0)\|$. In addition, PESp is also extended from global asymptotic stability in probability (Definition 3.1 in [78]). Notice that the stability notions in [77, 78] are global and PESp is local.

As for linear systems, one of the sufficient conditions for filter stability is uniform observability.

Definition 2.1.2 *[79] The pair (C_k, A_k) is uniformly observable if and only if*

there exist positive constants a, b, l*, for all* $k \geq 0$*, such that, for all* $k \geq 0$*,* $aI \leq \mathcal{M}_{k+l,k} \leq bI$ *where* $\mathcal{M}_{k+l,k} \triangleq \sum_{i=k}^{k+l} \Phi_{i,k} C_i C_i^T \Phi_{i,k}^T$ *is the observability gramian and* Φ_{k_1,k_0} *is the state transition matrix.*

Uniform observability reduces to observability if the linear system is time-invariant.

2.2 Motivating example

A power network is represented by undirected graph $(\mathcal{V}, \mathcal{E})$ with the set of buses $\mathcal{V} \triangleq \{1, \cdots, N\}$ and the set of transmission lines $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. The set of neighboring buses of $i \in \mathcal{V}$ is $\mathcal{S}_i \triangleq \{l \in \mathcal{V} \setminus \{i\} | (i,l) \in \mathcal{E}\}$. Each bus is either a generator bus $i \in \mathcal{G}$, or a load bus $i \in \mathcal{L}$. The dynamic of bus i with attacks is described as the following switched nonlinear system:

$$\begin{aligned} \dot{\theta}_i(t) &= f_i(t) + w_{1,i}(t) \\ \dot{f}_i(t) &= -\frac{1}{m_i}\Big(D_i f_i(t) + \sum_{l \in \mathcal{S}_i} P_{il}^{j_{il}(t)}(t) - (P_{M_i}(t) + d_{a,i}(t)) + P_{L_i}(t)\Big) + w_{2,i}(t) \end{aligned} \quad (2.1)$$

with the output model

$$y_{i,k} = [P_{elec_i,k}, \theta_{i,k}, f_{i,k}]^T + d_{s,i,k} + v_{i,k} \quad (2.2)$$

adopted from Chapter 9 in [80] adding a phase angle measurement as [81, 82]. System states $\theta_i(t)$, $f_i(t)$ are phase angle and angular frequency, respectively. Mode index $j_{il}(t) \in \{0, 1\}$ represents on/off of the power line connection between buses i and l; i.e., power flow is $P_{il}^1(t) = -P_{li}^1(t) = t_{il} \sin(\theta_i(t) - \theta_l(t))$, and $P_{il}^0(t) = -P_{li}^0(t) = 0$. The values $P_{L_i}(t)$, and $P_{elec_i}(t) = P_{L_i}(t) + D_i f_i(t)$ denote power demand, and electrical power output, respectively. Since power demand $P_{L_i}(t)$ can be obtained by many load forecasting methods [83, 84], it is assumed to be known.

Mechanical power $P_{M_i}(t)$ is the control input for $i \in \mathcal{G}$ and is assumed to be zero at load bus $i \in \mathcal{L}$. Power demand can be divided into elastic demand $P_{L_i}^E(t)$ and inelastic demand $P_{L_i}^{IE}(t)$ as shown in [85]; i.e., $P_{L_i}(t) = P_{L_i}^E(t) + P_{L_i}^{IE}(t)$. Elastic demand $P_{L_i}^E(t)$ can be controlled via power pricing. Since we assume that the current load is known, we simplify that load bus $i \in \mathcal{L}$ uses $P_{L_i}(t)$ as load controller.

The measurements are sampled at discrete instants due to hardware constraints. We use subscript $k \in \mathbb{Z}_{\geq 0}$ to denote an instantaneous value at the discrete sampling time t_k; e.g., $f_i(t_k) = f_{i,k}$.

An attacker is assumed to be able to modify the sensor measurements, control commands, and trigger the power flow line switches. The possible attacks are modeled as vectors $d_{s,i,k} \in \mathbb{R}$, $d_{a,i}(t) \in \mathbb{R}$, and hidden mode switch $j_{il}(t)$ which represent sensor attacks [31, 34], actuator attacks [31, 32, 69], and circuit breaking/switching attacks [36, 70], respectively.

2.3 Problem formulation

System model. Consider the hidden-mode nonlinear stochastic system

$$\begin{aligned} &\dot{x}(t) = f'(x(t), u(t) + d_a(t), w'(t), j(t), t), && x(t) \in \mathcal{C}^{j(t)} \\ &(x(t), j(t))^+ = \Omega'(x(t), j(t)), && x(t) \in \mathcal{D}^{j(t)} \\ &y_k = h(x_k, u_k + d_{a,k}, v'_k, j_k, t_k) + d_{s,k} \end{aligned} \tag{2.3}$$

where $x(t) \in \mathbb{R}^n$, $y_k \in \mathbb{R}^m$, $u(t) \in \mathbb{R}^s$, and $j(t) \in \mathbb{M}^I$ are state, output, input, and hidden-mode, respectively. We use subscript $k \in \mathbb{Z}_{\geq 0}$ to denote an instantaneous value at the discrete sampling time t_k. Vectors $d_a(t) \in \mathbb{R}^s$, and $d_{s,k} \in \mathbb{R}^m$ are actuator attack vector and sensor attack vector, respectively. Sets $\mathcal{C}^{j(t)}$, $\mathcal{D}^{j(t)} \subseteq \mathbb{R}^n$ denote flow set and jump set, respectively, and Ω is a mode transition function. For each mode, process noise $w'(t) \in \mathbb{R}^{s_1}$ and measurement noise $v'_k \in \mathbb{R}^{s_2}$ are uncorrelated with each other. The system is a continuous-discrete system because, while the physical dynamic evolves in continuous time, sensor measurements are obtained at their corresponding sampling instants due to hardware constraints. We define a uniform sampling period as $\epsilon = t_k - t_{k-1}$. It is assumed that the system (2.3) has a unique solution. One of the sufficient condition for the unique solution is weak one-sided local Lipschitz condition on function $f'(\cdot)$ in the open time interval of each mode duration [86] and other conditions can be found in the references therein. The system model (2.3) includes the power system model (2.1) with (2.2) as a special case.

Attack model. Signal attacks are comprised of signal magnitude attacks (i.e., the attacker injects attack signals), and signal location attacks (i.e., the attacker

chooses targeted sensors/actuators). Signal attacks are modeled by $d_a(t)$ and $d_{s,k}$ where zero values indicate that the corresponding actuators and sensors are free of attacks and non-zero values represent attack magnitudes. Switching attacks change system modes following Ω'.

Knowledge of the defender. The defender is unaware of which actuators/sensors are under attacks and what the current mode is. The defender knows dynamic system model and output model (2.4) for each mode but not the mode transition function Ω'. Mode set $\mathbb{M}^I$ is also known to the defender. The attack vectors $d_a(t)$, $d_{s,k}$, mode $j(t)$ and its transitions are inaccessible to the defender. Noise vectors $w'(t)$, v'_k are unknown but their auto covariance matrices are known.

Objective. The defender aims to answer the following three questions:
(a) if any sensor, actuator, or switch is attacked;
(b) if so, which ones are attacked, and how much sensor readings and control commands are tampered with;
(c) what current system states and mode are.
The above problem can be formulated as a joint estimation of states, attack vectors and modes of hidden-mode switched systems (2.3).

2.4 Estimator design

In order to reflect real world, system (2.3) models the attacks from the attacker's point of view and captures attack sources. In order to solve the estimation problem, we need to model the attacks from the defender's point of view and captures attack consequences. In particular, we rewrite system (2.3) as follows:

$$\begin{aligned} &\dot{x}(t) = f(x(t), u(t), d(t), w'(t), j(t), t), && x(t) \in \mathcal{C}^{j(t)} \\ &(x(t), j(t))^+ = \Omega(x(t), j(t)), && x(t) \in \mathcal{D}^{j(t)} \\ &y_k = h(x_k, u_k, v'_k, j_k, t_k) + H^{j_k} d_k \end{aligned} \tag{2.4}$$

where $d(t) = [d_a^T(t), d_s'^T(t)]^T \in \mathbb{R}^{s+m}$, $d'_{s,k} = h(x_k, u_k + d_{a,k}, v'_k, j_k, t_k) - h(x_k, u_k, v'_k, j_k, t_k) + d_{s,k}$, $f(x(t), u(t), d(t), w'(t), j(t), t) = f'(x(t), u(t) + S^{j(t)} d(t), w'(t), j(t), t)$, $S^{j(t)} = [K_S^{j(t)}, 0^{s\times m}] \in \{0,1\}^{s\times(s+m)}$ and $H^{j_k} = [0^{m\times s}, K_H^{j_k}] \in \{0,1\}^{m\times(s+m)}$. The

defender models the signal location attacks as mode $j(t)$ of diagonal matrix

$$K^j \triangleq \begin{bmatrix} S^j \\ H^j \end{bmatrix} = \begin{bmatrix} K_S^j & 0^{s\times m} \\ 0^{m\times s} & K_H^j \end{bmatrix} \in \{0,1\}^{(s+m)\times(s+m)}$$

where $K^j(i,i) = 1$ if mode j assumes that the i^{th} location is under attack; otherwise, $K^j(i,i) = 0$. Thus, $j(t) \in \mathbb{M} = \mathbb{M}^A \times \mathbb{M}^I$ stands for the both signal location attacks $\mathbb{M}^A$ and switching attacks $\mathbb{M}^I$.

Remark 2.4.1 *we will consider arbitrary $K^j \in R^{(s+m)\times(s+m)}$ in the remaining of this section and Sections 2.5, for the sake of generality.* ■

To solve the problem, we propose Nonlinear unknown Input, State and Mode Estimator (NISME). The NISME consists of a bank of Nonlinear unknown Input and State estimators (NISE) and a mode estimator as shown in Figure 2.1. Each

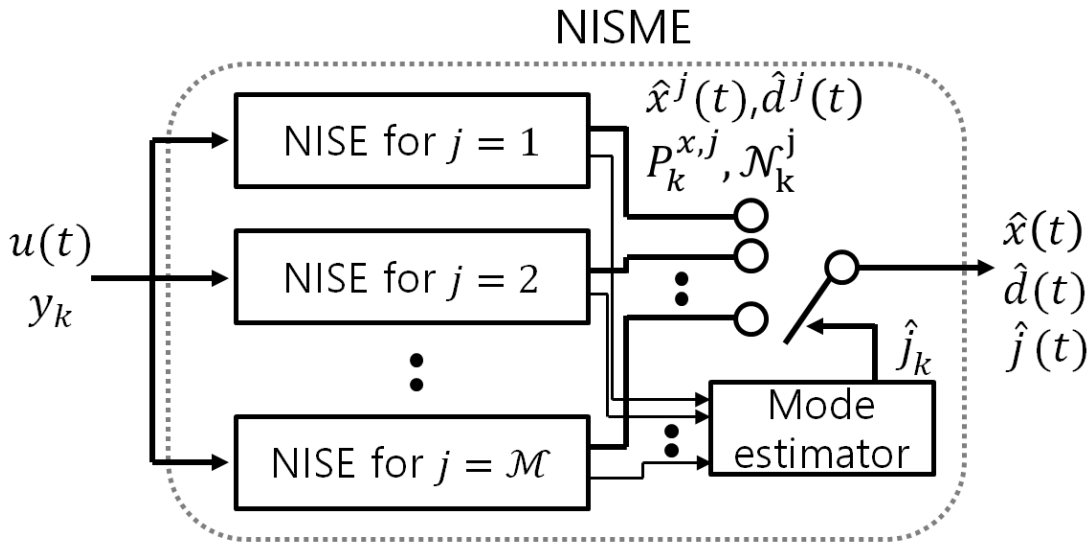

Figure 2.1: Scheme of NISME.

NISE is associated with a particular mode and recursively estimates the states and attack vectors under the fixed mode. The mode estimator calculates the posteriori probabilities of the modes by observing output discrepancies from predicted outputs, and chooses the most likely one. Lastly, the NISME outputs the estimates of the states and the attack vectors of the selected mode.

We first introduce some preliminaries for the NISE in Section 2.4.1. The NISME is presented in Section 2.4.2.

2.4.1 Preliminaries

In this section, we introduce an output decomposition used in the NISE. Since each NISE is associated with a particular mode, we omit the mode index $j(t)$ for notational simplicity.

We first discretize and linearize system (2.4) as follows with constant sampling period ϵ:

$$\begin{aligned} x_{k+1} &= x_k + \epsilon f(x_k, u_k, d_k, w'_k, t_k) + \epsilon\rho_k \\ &\simeq x_k + \epsilon(A_k x_k + B_k u_k + G_k d_k + \rho_k + w_k) \\ y_k &\simeq C_k x_k + D_k u_k + H d_k + v_k \end{aligned} \tag{2.5}$$

where $\epsilon\rho_k \triangleq \int_{t_k}^{t_{k+1}} f(x(\tau), u(\tau), d(\tau), w'(\tau), \tau)d\tau - \epsilon f(x_k, u_k, d_k, w'_k, t_k)$ refers to discretization error, and $w_k = J_k w'_k$, $v_k = E_k v'_k$,

$$\begin{aligned} A_k &\triangleq \frac{\partial f_k}{\partial x}\Big|_{\hat{x}_{k|k}, u_k, \hat{d}_k, 0, t_k}, & B_k &\triangleq \frac{\partial f_k}{\partial u}\Big|_{\hat{x}_{k|k}, u_k, \hat{d}_k, 0, t_k}, \\ G_k &\triangleq \frac{\partial f_k}{\partial d}\Big|_{\hat{x}_{k|k}, u_k, \hat{d}_k, 0, t_k}, & J_k &\triangleq \frac{\partial f_k}{\partial w'}\Big|_{\hat{x}_{k|k}, u_k, \hat{d}_k, 0, t_k}, \\ C_k &\triangleq \frac{\partial h_k}{\partial x}\Big|_{\hat{x}_{k|k-1}, u_k, 0, t_k}, & D_k &\triangleq \frac{\partial h_k}{\partial u}\Big|_{\hat{x}_{k|k-1}, u_k, 0, t_k}, \\ E_k &\triangleq \frac{\partial h_k}{\partial v'}\Big|_{\hat{x}_{k|k-1}, u_k, 0, t_k}. \end{aligned}$$

We define the autocovariance matrices for noise vectors as $\mathbb{E}[w_k w_k^T] = Q_k \geq 0$, and $\mathbb{E}[v_k v_k^T] = R_k > 0$.

Now we introduce two coordinate transformations. The first one is based on the singular value decomposition

$$H = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix}$$

where Σ is a full rank diagonal matrix. The first coordinate transformation T_k is defined by

$$T_k = \begin{bmatrix} T_{1,k} \\ T_2 \end{bmatrix} = \begin{bmatrix} I & -U_1^T R_k U_2 (U_2^T R_k U_2)^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix}. \tag{2.6}$$

Likewise, the singular value decomposition

$$T_2 C_k G_{k-1} V_2 = \begin{bmatrix} \bar{U}_{1,k} & \bar{U}_{2,k} \end{bmatrix} \begin{bmatrix} \bar{\Sigma}_k & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \bar{V}_{1,k}^T \\ \bar{V}_{2,k}^T \end{bmatrix}$$

with full-rank diagonal matrix $\bar{\Sigma}_k$ induces the second coordinate transformation

$$\bar{T}_k = [\bar{T}_{1,k}^T, \bar{T}_{2,k}^T]^T = \begin{bmatrix} I & -\bar{U}_{1,k}^T \bar{R}_k \bar{U}_{2,k} (\bar{U}_{2,k}^T \bar{R}_k \bar{U}_{2,k})^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} \bar{U}_{1,k}^T \\ \bar{U}_{2,k}^T \end{bmatrix} \tag{2.7}$$

where $\bar{R}_k \triangleq T_2 R_k T_2^T$. From coordinate transformations (2.6) and (2.7), the output y_k in (2.5) can be decomposed as follows:

$$\begin{aligned} y_{1,k} &= T_{1,k} y_k \simeq C_{1,k} x_k + D_{1,k} u_k + H_1 d_{1,k} + v_{1,k} \\ y_{2,k} &= \bar{T}_{1,k} T_2 y_k \simeq C_{2,k} x_k + D_{2,k} u_k + v_{2,k} \\ &\simeq C_{2,k}(x_{k-1} + \epsilon(A_{k-1} x_{k-1} + B_{k-1} u_{k-1} + G_{k-1} d_{k-1} + w_{k-1})) + D_{2,k} u_k + v_{2,k} \\ y_{3,k} &= \bar{T}_{2,k} T_2 y_k \simeq C_{3,k} x_k + D_{3,k} u_k + v_{3,k} \end{aligned} \tag{2.8}$$

where $H_{1,k} = \Sigma_k$ and $C_{2,k} G_{k-1} V_{2,k-1} \bar{V}_{1,k} = \bar{\Sigma}_k$. Note that $d_{1,k}$ and $d_{2,k}$ are different from $d'_{s,k}$ and $d_{a,k}$, and introduced for the purpose of analysis. Attack vector d_k is decomposed into a sum of $d_{1,k} \triangleq V_{1,k}^T d_k$, $d_{2,k} \triangleq \bar{V}_{1,k+1}^T V_{2,k}^T d_k$ and $d_{3,k} \triangleq \bar{V}_{2,k+1}^T V_{2,k}^T d_k$ where they are orthogonal to each other. In this case, it holds that $G_k d_k = G_{1,k} d_{1,k} + G_{2,k} d_{2,k} + G_{3,k} d_{3,k}$ with $G_{1,k} \triangleq G_k V_{1,k}$, $G_{2,k} \triangleq G_k V_{2,k} \bar{V}_{1,k+1}$ and $G_{3,k} \triangleq G_k V_{2,k} \bar{V}_{2,k+1}$.

Output $y_{1,k}$ is the portion of y_k which is attacked at k; i.e., $y_{1,k}$ includes $d_{1,k}$ in (2.8). Outputs $y_{2,k}$ and $y_{3,k}$ are the portions of y_k and are free of attacks at k, where output $y_{2,k}$ reflects $d_{2,k}$ indirectly because $C_{2,k} G_{k-1} d_{k-1} = \bar{\Sigma}$. Thus, decomposed outputs $y_{1,k}$, $y_{2,k}$, and $y_{3,k}$ are used to estimate $d_{1,k}$, $d_{2,k-1}$, and x_k, respectively.

Because $d_{2,k-1}$ is not measured by y_{k-1}, output $y_{2,k} = C_{2,k} G_{2,k-1} d_{2,k-1} + ...$ in (2.8) is instead used to estimate $d_{2,k-1}$; i.e., matrices $C_{2,k}$ and $G_{2,k-1}$ must be known to estimate attack vector $d_{2,k-1}$. However, in (2.5), matrix $G_{2,k-1}$ is obtained by linearizing $f(\cdot)$ using $\hat{d}_{2,k-1}$, and matrix C_k is obtained by linearizing $h(\cdot)$ using $\hat{x}_{k|k-1}$, where these linearizations cannot be done without knowing $\hat{d}_{2,k-1}$. Thus, we impose the following assumption on system (2.3) and its justification is given

in Section 2.6.

Assumption 2.4.1 *Dynamic system model* (2.4) *can be expressed as*

$$\begin{aligned}
\dot{x}(t) &= f(x(t), u(t), d_1(t), w'(t), t) + G_2(t)d_2(t) \\
y_{1,k} &= T_{1,k}y_k = h_1(x_k, u_k, v'_{1,k}, t_k) + H_1 d_{1,k} \\
y_{2,k} &= \bar{T}_{1,k}T_{2,k}y_k = C_{2,k}x_k + h_2(u_k, v'_{2,k}, t_k) \\
y_{3,k} &= \bar{T}_{2,k}T_{2,k}y_k = h_3(x_k, u_k, v'_{3,k}, t_k)
\end{aligned} \tag{2.9}$$

where $dim(d_{3,k}) = 0$.

With Assumption 2.4.1, the dynamic system (2.5) becomes

$$\begin{aligned}
x_{k+1} &= x_k + \epsilon f(x_k, u_k, d_{1,k}, w'_k, t_k) + \epsilon G_{2,k}d_{2,k} + \epsilon\rho_k \\
&\simeq x_k + \epsilon(A_k x_k + B_k u_k + G_{1,k}d_{1,k} + G_{2,k}d_{2,k} + \rho_k + w_k)
\end{aligned} \tag{2.10}$$

where matrices A_k, B_k, $G_{1,k}$, and J_k can be obtained before having an estimate for $d_{2,k}$. Output equation (2.9) is linearized into (2.8) where noises $v_{1,k} = E_{1,k}v'_{1,k}$, $v_{2,k} = E_{2,k}v'_{2,k}$, $v_{3,k} = E_{3,k}v'_{3,k}$ are uncorrelated with each other and

$$\begin{aligned}
C_{1,k} &\triangleq \frac{\partial h_{1,k}}{\partial x}\Big|_{\hat{x}_{k|k-1},u_k,0,t_k}, & C_{3,k} &\triangleq \frac{\partial h_{3,k}}{\partial x}\Big|_{\hat{x}_{k|k-1},u_k,0,t_k}, \\
D_{1,k} &\triangleq \frac{\partial h_{1,k}}{\partial u}\Big|_{\hat{x}_{k|k-1},u_k,0,t_k}, & D_{2,k} &\triangleq \frac{\partial h_{2,k}}{\partial u}\Big|_{u_k,0,t_k}, \\
D_{3,k} &\triangleq \frac{\partial h_{3,k}}{\partial u}\Big|_{\hat{x}_{k|k-1},u_k,0,t_k}, & E_{1,k} &\triangleq \frac{\partial h_{1,k}}{\partial v'_1}\Big|_{\hat{x}_{k|k-1},u_k,0,t_k}, \\
E_{2,k} &\triangleq \frac{\partial h_{2,k}}{\partial v'_2}\Big|_{u_k,0,t_k}, & E_{3,k} &\triangleq \frac{\partial h_{3,k}}{\partial v'_3}\Big|_{\hat{x}_{k|k-1},u_k,0,t_k}.
\end{aligned}$$

2.4.2 Algorithm statement

Consider the NISE (Algorithm 1) as well as Figure 2.2 whose derivation is presented in Appendix 2.8.1 in details. All the estimates of states and attack vectors are best linear unbiased estimates (BLUE); i.e., the estimator gains are chosen such that the estimates are unbiased and the norms of the error covariance matrices are minimized. Since attack vector $d_{2,k-1}$ does not influence output y_{k-1} directly, output $y^j_{2,k}$ is used to estimate attack vector $d_{2,k-1}$ (line 2) by using previous

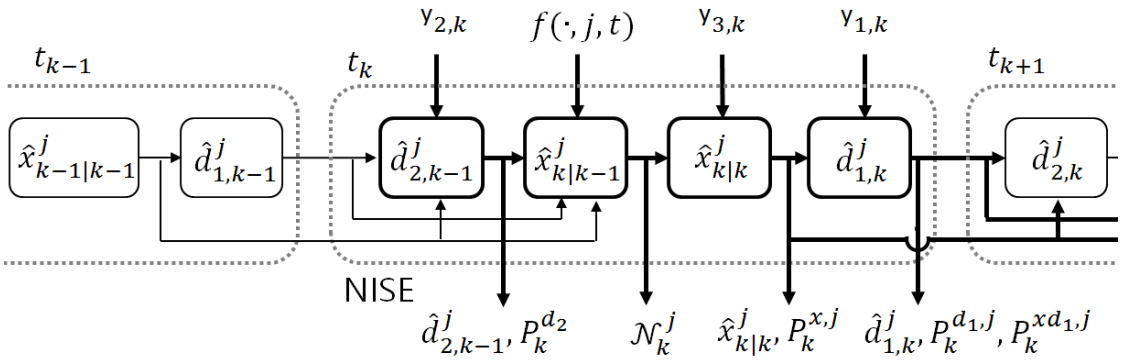

Figure 2.2: The recursive estimation scheme of the NISE.

estimate of attack vector $d_{1,k-1}$. Error covariance matrix $P^{d_2,j}_{k-1}$ of attack vector estimate $\hat{d}^j_{2,k-1}$ is derived in line 3. Applying the previous state and attack vector estimates to dynamic system (2.4), the current state is predicted (line 4). Error covariance matrix $P^{x,j}_{k|k-1}$ of the predicted state $\hat{x}^j_{k|k-1}$ is found in line 5 and the matrices in line 5 are defined by $\bar{Q}^j_{k-1} \triangleq \mathbb{E}[\bar{w}^j_{k-1}(\bar{w}^j_{k-1})^T]$,

$$\begin{aligned}
\bar{w}^j_{k-1} &\triangleq \epsilon(I - \epsilon G^j_{2,k-1} M^j_{2,k} C^j_{2,k})(w^j_{k-1} - G^j_{1,k-1} M^j_1 v^j_{1,k-1}) - \epsilon G^j_{2,k-1} M^j_{2,k} v^j_{2,k} \\
\bar{A}^j_{k-1} &\triangleq (I - \epsilon G^j_{2,k-1} M^j_{2,k} C^j_{2,k})(I + \epsilon A^j_{k-1} - \epsilon G^j_{1,k-1} M^j_1 C^j_{1,k-1}).
\end{aligned} \tag{2.11}$$

We correct the predicted state using the measurement bias (line 7) between the measured output and the predicted output. Error covariance matrix $P^{x,j}_{k|k-1}$ of state prediction $\hat{x}^j_{k|k-1}$ is updated in line 8. Attack vector $d_{1,k}$ is estimated from output $y^j_{1,k}$ (line 10). Error covariance matrix $P^{d_1,j}_{k-1}$ of attack vector estimate $\hat{d}^j_{1,k}$, and cross error covariance matrix $P^{xd_1,j}_k$ with $\hat{x}^j_{k|k}$ are found in lines 11-12. Lastly, the NISE generates the priori probability $\mathcal{N}^j_k$ (line 15) of the mode to find the most likely mode, where $n^j \triangleq Rank(\bar{P}^j_{k|k-1})$. For this purpose, the discrepancy between the measured output $y^j_{3,k}$ and the predicted output is used to validate the mode (line 13) because they should match if j is the true mode. Since the system is nonlinear, the discrepancy ν^j_k may not be Gaussian. We approximate ν^j_k as a Gaussian random vector because it is a typical practice to approximate an unknown noise as a Gaussian distribution as [87]. Moreover, ν^j_k is Gaussian when the system is linear and noises w^j_k and v^j_k are Gaussian. Covariance matrix $\bar{P}^j_{k|k-1}$ of the discrepancy ν^j_k is found in line 14.

Algorithm 1 NISE

Input: $j, \hat{x}^j_{k-1|k-1}, \hat{d}^j_{1,k-1}, P^{x,j}_{k-1}, P^{d_1,j}_{k-1}, P^{xd_1,j}_{k-1}, y_k, u(t)$ for $t \in [t_{k-1}, t_k]$;

1: ▷ ***Attack vector $\boldsymbol{d}^j_{2,k-1}$ estimation***

2: $M^j_{2,k} = (\epsilon C^j_{2,k} G^j_{2,k-1})^{-1}$ (or $M^j_{2,k} = 0$ if $Rank(\bar{\Sigma}_k) = 0$);

3: $\hat{d}^j_{2,k-1} = M^j_{2,k}(y^j_{2,k} - C^j_{2,k}(\hat{x}^j_{k-1|k-1} + \epsilon f(\hat{x}^j_{k-1|k-1}, u_{k-1}, \hat{d}^j_{1,k-1}, 0, j, t_{k-1})) - h_2(u_k, 0, j, t_k))$;

4: $P^{d_2,j}_{k-1} = M^j_{2,k} C^j_{2,k}(I + \epsilon A^j_{k-1}) P^j_{k-1} (M^j_{2,k} C^j_{2,k}(I + \epsilon A^j_{k-1}))^T + \epsilon^2 M^j_{2,k} C^j_{2,k} Q^j_{k-1} (M^j_{2,k} C^j_{2,k})^T + M^j_{2,k} R^j_{2,k} (M^j_{2,k})^T + \epsilon^2 C^j_{2,k} G^j_{1,k-1} P^{d_1,j}_{k-1} (C^j_{2,k} G^j_{1,k-1})^T + \epsilon M^j_{2,k} C^j_{2,k}(I + \epsilon A^j_{k-1}) P^{xd_1,j}_{k-1} (C^j_{2,k} G^j_{1,k-1})^T + \epsilon C^j_{2,k} G^j_{1,k-1} P^{d_1 x,j}_{k-1} (M^j_{2,k} C^j_{2,k}(I + \epsilon A^j_{k-1}))^T$;

5: ▷ ***State prediction***

6: $\dot{\hat{x}}^j(t) = f(\hat{x}^j(t), u(t), \hat{d}^j_{1,k-1}, 0, j, t) + G^j_{2,k-1} \hat{d}^j_{2,k-1}$ with initial condition $\hat{x}^j_{k-1|k-1}$ for $t \in (t_{k-1}, t_k]$ to have $\hat{x}^j_{k|k-1}$ at $t = t_k$;

7: $P^{x,j}_{k|k-1} = \bar{A}^j_{k-1} P^{x,j}_{k-1} (\bar{A}^j_{k-1})^T + \bar{Q}^j_{k-1}$;

8: ▷ ***State estimation***

9: $L^j_k = P^{x,j}_{k|k-1} (C^j_{3,k})^T (C^j_{3,k} P^{x,j}_{k|k-1} (C^j_{3,k})^T + R^j_{3,k})^{-1}$;

10: $\hat{x}^j_{k|k} = \hat{x}^j_{k|k-1} + L^j_k (y^j_{3,k} - h_3(\hat{x}^j_{k|k-1}, u_k, 0, j, t_k))$;

11: $P^{x,j}_k = (I - L^j_k C^j_{3,k}) P^{x,j}_{k|k-1} (I - L^j_k C^j_{3,k})^T + L^j_k R^j_{3,k} (L^j_k)^T$;

12: ▷ ***Attack vector $d^j_{1,k}$ estimation***

13: $M^j_1 = (H^j_1)^{-1}$ (or $M^j_1 = 0$ if $Rank(\Sigma) = 0$);

14: $\hat{d}^j_{1,k} = M^j_1 (y^j_{1,k} - h_1(\hat{x}^j_{k|k}, u_k, 0, j, t_k))$;

15: $P^{d_1,j}_k = M^j_1 C^j_{1,k} P^{x,j}_k (M^j_1 C^j_{1,k})^T + M^j_1 R^j_{1,k} (M^j_1)^T$;

16: $P^{xd_1,j}_k = -P^{x,j}_k (M^j_1 C^j_{1,k})^T$;

17: ▷ ***The priori probability of the mode***

18: $\nu^j_k = y^j_{3,k} - h_3(\hat{x}^j_{k|k-1}, u_k, 0, j, t_k)$;

19: $\bar{P}^j_{k|k-1} = C^j_{3,k} P^j_{k|k-1} (C^j_{3,k})^T + R^j_{3,k}$;

20: $\mathcal{N}^j_k = \frac{1}{(2\pi)^{n^j/2} |\bar{P}^j_{k|k-1}|^{1/2}} \exp(-\frac{(\nu^j_k)^T (\bar{P}^j_{k|k-1})^{-1} \nu^j_k}{2})$;

Return: $\hat{x}^j_{k|k}, \hat{d}^j_{1,k}, \hat{d}^j_{2,k-1}, P^{x,j}_k, P^{d_2,j}_{k-1}, P^{d_1,j}_k, P^{xd_1,j}_k, \mathcal{N}^j_k$.

Algorithm 2 NISME

Input: $\hat{x}^j_{0|0} = \mathbb{E}[x_0]$, $P^{x,j}_0 = P_0$, $\mu^j_0 = \frac{1}{|\mathbb{M}|}$, $\hat{d}^j_{1,0} = (\Sigma^j)^{-1}(y^j_{1,0} - h_1(\hat{x}^j_{0|0}, u_0, 0, j, 0))$ for $\forall j \in \mathbb{M}$; Choose $0 < \delta \ll \frac{1}{|\mathbb{M}|}$, $0 < \alpha_1 < 1$, $0 < \alpha_2 < 1$ (significance levels);

for $k = 1 : N$ **do**

  Read sensor output y_k, and control input $u(t)$ for $t \in [t_{k-1}, t_k]$;

  for $j \in \mathbb{M}$ **do**

    Run the NISE with input $(j, \hat{x}^j_{k-1|k-1}, \hat{d}^j_{1,k-1}, P^{x,j}_{k-1}, P^{d_1,j}_{k-1}, P^{xd_1,j}_{k-1}, y_k, u(t)$ for $t \in [t_{k-1}, t_k])$ to generate output $(\hat{x}^j_{k|k}, \hat{d}^j_{1,k}, \hat{d}^j_{2,k-1}, P^{x,j}_k, P^{d_2,j}_{k-1}, P^{d_1,j}_k, P^{xd_1,j}_k, \mathcal{N}^j_k)$;

  end for

▷ ***Mode estimator***

  for $j \in \mathbb{M}$ **do**

    $\bar{\mu}^j_k = \max\{\frac{\mathcal{N}^j_k \mu^j_{k-1}}{\sum_{i=1}^{|\mathbb{M}|} \mathcal{N}^i_k \mu^i_{k-1}}, \delta\}$;

  end for

  for $j \in \mathbb{M}$ **do**

    $\mu^j_k = \frac{\bar{\mu}^j_k}{\sum_{i=1}^{|\mathbb{M}|} \bar{\mu}^i_k}$;

  end for

  Set $\hat{j}_k = \text{argmax}_j\, \mu^j_k$;

  Obtain $\chi^2_{|\hat{d}^{\hat{j}_k}_{1,k}|}(\alpha_1)$ and $\chi^2_{|\hat{d}^{\hat{j}_k}_{2,k-1}|}(\alpha_2)$;

  if $(\hat{d}^{\hat{j}_k}_{1,k})^T (P^{d_1,\hat{j}_k}_k)^{-1} \hat{d}^{\hat{j}_k}_{1,k} < \chi^2_{|\hat{d}^{\hat{j}_k}_{1,k}|}(\alpha_1)$ **and** $(\hat{d}^{\hat{j}_k}_{2,k-1})^T (P^{d_2,\hat{j}_k}_{k-1})^{-1} \hat{d}^{\hat{j}_k}_{2,k-1} < \chi^2_{|\hat{d}^{\hat{j}_k}_{2,k-1}|}(\alpha_2)$ **then**

    Set $\hat{j}^{true}_k$ as signal attack-free mode of $\hat{j}_k$;

    $\hat{d}^{\hat{j}_k}_{1,k} = \hat{d}^{\hat{j}_k}_{2,k-1} = 0$;

  end if

  Return:

$$\hat{j}(t) = \hat{j}^{true}_k \text{ for } t \in (t_{k-1}, t_k],\ \hat{x}(t) = \hat{x}^{\hat{j}_k}_{k|k}, t \in (t_{k-1}, t_k],$$
$$\hat{d}_1(t) = \hat{d}^{\hat{j}_k}_{1,k} \text{ for } t \in [t_k, t_{k+1}),\ \hat{d}_2(t) = \hat{d}^{\hat{j}_k}_{2,k-1} \text{ for } t \in (t_{k-1}, t_k].$$

end for

Now consider the NISME (Algorithm 2) which is derived in Section 2.8.2. The NISME runs the NISE for each mode $j \in \mathbb{M}$ in parallel to generate the state

and attack vector estimates along with the priori probability for each mode (line 4). After then, the algorithm identifies the most likely mode (lines 6-11). By the Bayes' theorem, the posteriori probability μ_k^j is updated by a linear combination of the priori probabilities (line 7). It is not desirable that some mode probabilities vanish over time because the true modes might be time-varying. A lower bound δ is adopted in line 7 to prevent the vanishment of the mode probabilities. After the lower bound is applied, the mode probability is normalized in line 10. The mode with the largest posteriori probability μ_k^j is chosen as a current mode (line 12), and the attack vectors of the current mode are tested by Chi-square hypothesis tests (p.354 in [88]) with significance levels α_1, α_2 to determine whether they are statistically significant or not (line 14). Specifically, we have the following null-hypothesis and alternative hypothesis

$$\mathcal{H}_0 : d_{1,k} = 0 \text{ and } d_{2,k-1} = 0, \quad \mathcal{H}_1 : d_{1,k} \neq 0 \text{ or } d_{2,k-1} \neq 0$$

with samples $\hat{d}_{1,k}^{\hat{j}_k}$ and $\hat{d}_{2,k-1}^{\hat{j}_k}$. Chi-square value is presented as $\chi_{df}^2(\alpha)$ where df and α are the degree of freedom and significance level, respectively. If it is not statistically significant, the algorithm chooses the signal attack-free mode as a current mode. The corresponding state and attack vector estimates are returned (line 18). Due to limited measurements over the continuous-time dynamic system model, we use the approximation that the attack vector estimates are constants during a sampling period, in lines 2,4,10 of the NISE, and lines 18 of the NISME. We, however, will consider approximation errors in the analysis.

2.5 Analysis

We consider the linearization errors ϕ_k, $\psi_{1,k}$, $\psi_{2,k}$, and $\psi_{3,k}$ defined by

$$\begin{aligned}
&f(x_k, u_k, d_{1,k}, w_k', t_k) - f(\hat{x}_{k|k}, u_k, \hat{d}_{1,k}, 0, t_k) \\
&= A_k \tilde{x}_{k|k} + G_{1,k} \tilde{d}_{1,k} + w_k + \phi_k(\hat{x}_{k|k}, x_k, u_k, w_k', v_k') \\
&h_1(x_k, u_k, v_k', t_k) - h_1(\hat{x}_{k|k}, u_k, 0, t_k) = C_{1,k} \tilde{x}_{k|k} + v_{1,k} + \psi_{1,k}(\hat{x}_{k|k}, x_k, u_k, v_k') \\
&h_2(u_k, v_k', t_k) - h_2(u_k, 0, t_k) = v_{2,k} + \psi_{2,k}(u_k, v_k') \\
&h_3(x_k, u_k, v_k', t_k) - h_3(\hat{x}_{k|k-1}, u_k, 0, t_k) = C_{3,k} \tilde{x}_{k|k-1} + v_{3,k} + \psi_{3,k}(\hat{x}_{k|k-1}, x_k, u_k, v_k')
\end{aligned}$$

where ϕ_k is a function of $\tilde{d}_{1,k}$ and $\tilde{d}_{1,k}$ is a function of x_k, $\hat{x}_{k|k}$ and v'_k. We omit the arguments of the linearization errors in the rest of the paper for notational simplicity. The following set of assumptions is needed to ensure the stability of the NISE algorithm, and justified in Section 2.6.1.

Assumption 2.5.1 *Attack vector $d(t)$ is continuous, and its slopes are uniformly bounded; i.e., there exists $\bar{d} > 0$ such that $\sup_{t_1,t_2\geq 0} \|(d(t_1) - d(t_2))/(t_1 - t_2)\| \leq \bar{d}$.*

Assumption 2.5.2 *There exist $\bar{a}'$, $\bar{c}_3$, $\underline{q}'$, $\underline{r}_3 > 0$ such that the following holds for $k \geq 0$:*

$$\|A_k\| \leq \bar{a}',\ \|C_{3,k}\| \leq \bar{c}_3,\ \underline{q}' \leq Q_k,\ \underline{r}_3 I \leq R_{3,k}.$$

If $rk(\Sigma_k) \neq 0$, there exist $\bar{c}_1$, $\bar{g}_1$, $\bar{m}_1 > 0$ such that the following holds for $k \geq 0$:

$$\|C_{1,k}\| \leq \bar{c}_1,\ \|G_{1,k}\| \leq \bar{g}_1,\ \|\Sigma_k^{-1}\| \leq \bar{m}_1.$$

If $rk(\bar{\Sigma}_k) \neq 0$, there exist $\bar{c}_2$, $\underline{g}_2$, $\bar{g}_2$, $\underline{m}_2$, $\bar{m}_2$, $\underline{r}_2 > 0$ such that the following holds for $k \geq 0$:

$$\|C_{2,k}\| \leq \bar{c}_2,\ \underline{g}_2 \leq \|G_{2,k}\| \leq \bar{g}_2,\ \underline{m}_2 \leq \|\bar{\Sigma}_k^{-1}\| \leq \bar{m}_2,\ \underline{r}_2 I \leq R_{2,k}.$$

Assumption 2.5.3 *For any $\epsilon_\phi, \epsilon_{\psi_1}, \epsilon_{\psi_2}, \epsilon_{\psi_3} > 0$, there exists $\delta > 0$ such that*

$$\|\phi_k\| \leq \epsilon_\phi \|x_k - \hat{x}_{k|k}\|^2, \quad \|\psi_{1,k}\| \leq \epsilon_{\psi_1} \|x_k - \hat{x}_{k|k}\|^2$$
$$\|\psi_{2,k}\| \leq \epsilon_{\psi_2} \|x_k - \hat{x}_{k|k}\|^2, \quad \|\psi_{3,k}\| \leq \epsilon_{\psi_3} \|x_k - \hat{x}_{k|k}\|^2$$

hold for all $\|x_k - \hat{x}_{k|k}\| \leq \delta$ and $k \geq 0$.

Assumption 2.5.4 *There exist $\underline{p}$, $\bar{p} > 0$ such that $\underline{p} I \leq P_k^x \leq \bar{p} I$ for $k \geq 0$.*

Let us denote the discretization error for the state prediction as

$$\hat{\rho}_k \triangleq \int_{t_k}^{t_{k+1}} f(\hat{x}(\tau), u(\tau), \hat{d}_{1,k}, 0, \tau) d\tau - \epsilon f(\hat{x}_{k|k}, u_k, \hat{d}_{1,k}, 0, t_k).$$

Assumption 2.5.5 *There exist $\epsilon_\rho, \delta_\rho > 0$ such that $\|\rho_k\| \leq \epsilon^2 \epsilon_\rho$ and $\|\hat{\rho}_k\| \leq \epsilon^2 \epsilon_\rho$ for all $\epsilon \leq \delta_\rho$ and $k \geq 0$, where ρ_k is defined in (2.5).*

Under the above assumptions, the following theorem ensures PESp-like properties for the estimation errors.

Theorem 2.5.1 *Consider the NISE algorithm, provided that Assumptions 2.4.1, 2.5.1, 2.5.2, 2.5.3, 2.5.4 and 2.5.5 hold. For any $\gamma \in (0,1)$, there exists a set of positive constants α_x, α_{d_1}, α_{d_2}, b_x, b_{d_1}, b_{d_2}, c_x, c_{d_1}, c_{d_2}, $\underline{\delta}$, $\bar{q}'$, $\bar{r}_1$, $\bar{r}_2$, $\bar{r}_3$, and $\bar{\epsilon}$ such that, if $Q_k \leq \bar{q}' I$, $R_{1,k} \leq \bar{r}_1 I$, $R_{2,k} \leq \bar{r}_2 I$, $R_{3,k} \leq \bar{r}_3 I$, and $\epsilon \leq \bar{\epsilon}$, then the following properties hold:*

$$
\begin{aligned}
P(\|\tilde{x}_{k|k}\| < \alpha_x e^{-b_x k}\|\tilde{x}_{0|0}\| + c_x) \geq 1-\gamma,\\
P(\|\tilde{d}_1(t)\| < \alpha_{d_1} e^{-b_{d_1} t}\|\tilde{x}_{0|0}\| + c_{d_1}) \geq 1-\gamma,\\
P(\|\tilde{d}_2(t)\| < \alpha_{d_2} e^{-b_{d_2} t}\|\tilde{x}_{0|0}\| + c_{d_2}) \geq 1-\gamma
\end{aligned}
$$

for all $\|\tilde{x}_{0|0}\| \leq \underline{\delta}$, $k \geq 0$ and $t \geq 0$.

Theorem 2.5.1 is formally proven in Appendix 2.8.3.

Remark 2.5.1 *The set of constants in Theorem 2.5.1 can be obtained in the proof. For ease of presentation, we omit the procedure to find these constants.* ■

2.6 Discussion

2.6.1 Assumption justification

The proposed NISE algorithm is an extension of ISE for linear systems in [89–91] to nonlinear systems. It is also an extension of the EKF in [92–95] to include unknown inputs. Thus, the assumptions in Theorem 2.5.1 are similar to the assumptions therein.

In Assumption 2.4.1, $dim(d_{3,k}) = 0$ (or equivalently, $rk(C_{2,k}G_{2,k-1}) = p - dim(d_{1,k-1})$) is a required condition for the stability of ISE for linear systems (Theorem 5 in [89]). Assumption 2.4.1 also requires that the system function is partially linear. Its justification is given below. Since $dim(d_{3,k}) = 0$, unknown input d_{k-1} is decomposed into two orthogonal vectors $d_{1,k-1}$ and $d_{2,k-1}$. From (2.8), one can see that $d_{1,k-1}$ is included in $y_{1,k-1}$ and thus y_{k-1}, but $d_{2,k-1}$ is not. Then one needs to estimate $d_{2,k-1}$ from $y_{2,k}$. In particular, $y_{2,k}$ includes the term $C_{2,k}G_{k-1}d_{k-1}$ which can be decomposed as $C_{2,k}G_{k-1}d_{k-1} = C_{2,k}G_{1,k-1}d_{1,k-1} + C_{2,k}G_{2,k-1}d_{2,k-1}$.

However, in system (2.5), we obtain matrix $G_{2,k-1}$ by linearizing f' using $\hat{d}_{2,k-1}$ and transforming G_{k-1} by (2.6) and (2.7). We get matrix $C_{2,k}$ by linearizing h using $\hat{x}_{k|k-1}$ and transforming C_k by (2.6) and (2.7), where state prediction $\hat{x}_{k|k-1}$ needs unknown input estimate $\hat{d}_{2,k-1}$ in line 7. These linearizations cannot be done without $\hat{d}_{2,k-1}$. Thus, Assumption 2.4.1 requires that nonlinear function f is independent of $d_2(t)$ and the function including $d_2(t)$ is linear. Furthermore, Assumption 2.4.1 is satisfied either H_k in system (2.3) has full-column rank, or system (2.3) is in the following form:

$$\dot{x}(t) = f(x(t), u(t), w'(t), t) + G(t)d(t)$$
$$y_k = C_k x_k + h'(u_k, v'_k, t_k) + H_k d_k.$$

A sufficient condition for Assumption 2.5.1 is Assumption 1 in [90] where $d(t)$ is assumed to be differentiable for all t and its gradients are uniformly bounded.

Assumptions 2.5.2, 2.5.3 and 2.5.4 reduce to Assumption 3.1 in [93] when system (2.3) becomes a discrete time system; i.e., $\rho_k = \hat{\rho}_k = 0$, and unknown input $d(t)$ is absent. A sufficient condition for Assumptions 2.5.2 is that functions f' and h satisfy bi-Lipschitz continuity (p.10 in [96]), and covariance matrices of noise vectors w'_k and v'_k are uniformly lower bounded by positive definite matrices. Assumptions 2.5.3 can be verified if functions f' and h satisfy Holder continuity (p.136 in [97]) with exponent 2. Uniform observability of the pair $(C_{3,k}, \bar{A}_k)$ is a sufficient condition for Assumption 2.5.4, as shown in Lemma 2.6.1. Recall that $C_{3,k}$ and $\bar{A}_k$ are defined in (2.8) and (2.11).

Lemma 2.6.1 *Consider the NISE algorithm. Under Assumptions 2.4.1, 2.5.2, and 2.5.3, if the pair $(C_{3,k}, \bar{A}_k)$ is uniformly observable and $P_0^x \geq 0$, there exist $\underline{p}, \bar{p} > 0$ such that $\underline{p}I \leq P_k^x \leq \bar{p}I$ for all $k \geq 0$. Moreover, if there exist $\bar{q}'$, $\bar{r}_1$, $\bar{r}_2 \geq 0$ such that $Q_k \leq \bar{q}'I$, $R_{1,k} \leq \bar{r}_1 I$, and $R_{2,k} \leq \bar{r}_2 I$, then there exist $\bar{p}^{d_1}, \bar{p}^{d_2} > 0$ such that $0 \leq P_k^{d_1} \leq \bar{p}^{d_1} I$, $0 \leq P_k^{d_2} \leq \bar{p}^{d_2} I$ for all $k \geq 0$.*

Lemma 2.6.1 is proven in Appendix 2.8.3. Lemma 2.6.1 does not require Assumptions 2.5.1 and 2.5.5 because the update law of error covariance matrices depends only on known system matrices and covariance matrices of the noises.

Assumption 2.5.5 is fulfilled if f' is Lipschitz and bounded, and $u(t), d(t), w'(t)$ are Lipschitz in the small (Definition 2.14 in [98]).

2.6.2 Connections to existing results

Connections to the EKF. If system (2.3) is a discrete time system and unknown input $d(t)$ is absent, then the NISE algorithm reduces to the EKF [93, 94, 99, 100].

Connections to the ISE for linear systems. If system (2.3) is a discrete time linear system and unknown input $d_2(t)$ is absent, then the NISE algorithm reduces to the filter in [89]. As the EKF, we design the NISE algorithm such that the estimates are optimal as if system (2.3) is linear. To be specific, let $dim(d_{3,k}) = 0$ and initial estimate $\hat{x}_{0|0}$ be unbiased. Then, state estimate $\hat{x}_{k|k}$ and unknown input estimates $\hat{d}_{1,k}$, $\hat{d}_{2,k}$ are the minimum variance unbiased estimates among all the linear estimates, if functions f' and h are linear. The proof is similar to that of Lemma 3 in [89].

2.6.3 Mode reduction

When there are $s + m$ signal attack locations, mode set $\mathbb{M}^A$ includes all the combinations of the attack locations; i.e., $|\mathbb{M}^A| = 2^{s+m}$. As the number of signal attack location increases, computational complexity increases exponentially. We, in this section, discuss how to alleviate computational complexity by reducing the number of modes induced by $\mathbb{M}^A$, and how to estimate the true mode from the estimation results of a reduced mode set for hidden-mode switched linear systems. Finding a reduced mode set presented in this section, and finding a true mode presented in Section 2.6.4 are both on-line procedures. The former is conducted before running the NISE, and the latter will replace the hypothesis test (line 13-17) in the NISME. The complete algorithm for the NISME with reduced mode set is presented in Section 2.6.4, where we consider the special case with $|\mathbb{M}^I| = 1$. In Section 2.6.5, we extend the results to any $\mathbb{M}^I$.

Let us consider the special case where the dynamic for each mode j is linear and switching, and $|\mathbb{M}^I| = 1$:

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)(u(t) + S^{j(t)}d(t)) + w(t), && x(t) \in \mathcal{C}^{j(t)} \\ (x(t), j(t))^+ &= \Omega(x(t), j(t)), && x(t) \in \mathcal{D}^{j(t)} \\ y_k &= C_k x_k + D_k u_k + H^{j_k} d_k + v_k \end{aligned} \tag{2.12}$$

where $S^{j(t)}$, and H^{j_k} are defined in (2.4). We remind $K^j = [(S^j)^T, (H^j)^T]^T$, and

define $\mathbb{K}_1^j \triangleq \{i = 1, \cdots, n+m | K^j(i,i) = 1\}$.

Mode reduction is based on the following two ideas. Firstly, we maintain the modes such that uniform observability over finite time-horizon $[c_1, c_2]$ (Definition 2.6.1) holds. Secondly, we remove modes whose assumptions on attack locations are strictly restrictive than those of others.

Definition 2.6.1 *The pair (C_k, A_k) is uniformly observable over $[c_1, c_2]$ if and only if there exist positive constants a, b, and $l < c_2 - c_1$ such that $aI \leq \mathcal{M}_{k+l,k} \leq bI$ for $k = c_1, c_1 + 1, \cdots, c_2 - l$.*

The first idea is motivated by the sufficient condition for Lemma 2.6.1; uniform observability. To check uniform observability, the defender is required to have information on pairs $(C_{3,k}^j, \bar{A}_k^j)$ for $k = 0, 1, \cdots$. This information is hard to gather at initial time. We instead adopt an approximation, Definition 2.6.1, which only requires the system matrices for a few next steps. Uniform observability over $[c_1, c_2]$ reduces to uniform observability as $c_2 \to \infty$ with fixed c_1.

To justify the second idea, consider a pair of modes $j, j' \in \mathbb{M}^A$ such that $\mathbb{K}_1^{j'} \subset \mathbb{K}_1^j$, and $(C_{3,k}^j, \bar{A}_k^j)$, $(C_{3,k}^{j'}, \bar{A}_k^{j'})$ are uniformly observable over $[c_1, c_2]$. The relation $\mathbb{K}_1^{j'} \subset \mathbb{K}_1^j$ indicates that mode j' imposes a more restrictive assumption on attack locations than mode j. In this sense, mode j' is said to be redundant and it could be ruled out to reduce computational complexity.

Intuitively speaking, the above ideas allow the minimal number of modes to provide the same attack capability as the power set. The reduced mode set is defined by $\mathbb{M}_{[c_1,c_2]}^d = \{j \in \mathbb{M}_{[c_1,c_2]}^{ob} | \nexists j' \in \mathbb{M}_{[c_1,c_2]}^{ob} \text{ s.t. } \mathbb{K}_1^j \subset \mathbb{K}_1^{j'}\}$ where $\mathbb{M}_{[c_1,c_2]}^{ob} \triangleq \{j \in \mathbb{M}^A | (C_{3,k}^j, \bar{A}_k^j)$ is uniformly observable over $[c_1, c_2]\}$. It can be found by Algorithm 3 where $\mathbb{M}_i^A \triangleq \{j \in \mathbb{M}^A | |\mathbb{K}_1^j| = i\}$. Without the mode reduction, the worst upper bound of $|\mathbb{M}^A|$ is 2^{s+m}, but $\mathbb{M}_{[c_1,c_2]}^d$ could be as low as 1 (see case study 2). It is worthy to emphasize that the defender needs to know $(C_{3,k}^j, \bar{A}_k^j)$ over $[c_1, c_2]$ in order to verify uniform observability in the interval.

Algorithm 3 Mode reduction (finding $\mathbb{M}^d_{[c_1,c_2]}$).

Input: $\mathbb{M}^A$, A_k, C_k for $k = c_1, c_1+1, \cdots, c_2$ (or corresponding $A(t)$, $C(t)$), G^j, H^j for $\forall j \in \mathbb{M}^A$;

1: $\mathbb{M}^d_{[c_1,c_2]} = \emptyset$;
2: **for** $i = s+m:1$ **do**
3: **for** $j \in \mathbb{M}^A_i$ **do**
4: **if** $(C^j_{3,k}, \bar{A}^j_k)$ is uniformly observable over $[c_1, c_2]$, **and** $\nexists j' \in \mathbb{M}^d_{[c_1,c_2]}$ s.t. $\mathbb{K}^j_1 \subset \mathbb{K}^{j'}_1$ **then**
5: $\mathbb{M}^d_{[c_1,c_2]} = \mathbb{M}^d_{[c_1,c_2]} \cup \{j\}$;
6: **end if**
7: **end for**
8: **end for**

Return: $\mathbb{M}^d_{[c_1,c_2]}$.

Remark 2.6.1 *Mode reduction Algorithm 3 is not applicable to nonlinear systems because uniform observability is determined by where linearization is performed. This information cannot be obtained in advance.* ■

2.6.4 True mode estimation

We discuss how to estimate the true mode from the outputs of the NISME under the reduced mode set. It might be noticed that the reduced mode set might not include true modes, since some of the modes are removed. Given mode estimate $\hat{j}(t)$ from the reduced mode set $\mathbb{M}^d_{[c_1,c_2]}$, the idea is to conduct two-tailed z-test [88] for each attack location $i \in \mathbb{K}^{\hat{j}(t)}_1$ to determine whether the attack size is statistically significant. To be specific, we test the null hypothesis that i^{th} elements of $d_{1,k}$ or $d_{2,k-1}$ are zero:

$$\mathcal{H}_0 : d_{1,k}(i-s) = 0 \text{ if } i > s,\ \mathcal{H}_0 : d_{2,k-1}(i) = 0 \text{ if } i \leq s$$

and $\hat{d}^{\hat{j}_k}_{1,k}(i-s)$ or $\hat{d}^{\hat{j}_k}_{2,k-1}(i)$ are regarded as samples. If the null hypothesis is rejected, then we accept alternative hypothesis

$$\mathcal{H}_1 : d_{1,k}(i-s) \neq 0 \text{ if } i > s,\ \mathcal{H}_1 : d_{2,k-1}(i) \neq 0 \text{ if } i \leq s$$

i.e., there exists an attack on i^{th} location. Algorithm 4 presents the pseudo code for true mode estimation. z-value is presented as $z(\alpha)$ where α is the significance level. Hypothesis tests are conducted in lines 6 and 13 for actuator attacks, and sensor attacks, respectively.

2.6.5 NISME with reduced mode set

Algorithm 5 shows the NISME with reduced mode sets. The core of the algorithm is identical to that of the NISME, and some differences are explained as follows. We apply the mode reduction technique and true mode estimation technique to each $i \in \mathbb{M}^I$. The algorithm calculates reduced mode set every $\mathcal{T}$ steps for every $i \in \mathbb{M}^I$ (line 4). This requires the defender to have knowledge on system matices for next $\mathcal{T}-1$ steps. Based on the fact that the reduced mode set might not include the true mode, we test attack vectors element-wise to identify the true mode (line 19). As $\mathcal{T}$ decreases in Algorithm 5, lesser knowledge on system matrices is required, but computational complexity induced by Algorithm 3 increases. When system (2.12) is time-invariant, $\mathcal{T} = \infty$.

Algorithm 4 Mode estimation with mode reduction

Input: $\hat{j}_k$, $\hat{d}^{\hat{j}_k}_{1,k}$, $\hat{d}^{\hat{j}_k}_{2,k-1}$, $P^{d_1,\hat{j}_k}_k$, $P^{d_2,\hat{j}_k}_{k-1}$, α_1, α_2 (significance levels);

Obtain z-values $z(\alpha_1)$ and $z(\alpha_2)$ from z-test table;
$K^{\hat{j}^{true}_k} = 0^{(s+m)\times(s+m)}$;
$l_1 = l_2 = 1$;
for $i \in \mathbb{K}_1^{\hat{j}_k}$ **do**
  if $i \leq s$ **then**
    if $\frac{|\hat{d}^{\hat{j}_k}_{2,k-1}(l_2)|}{\sqrt{P^{d_2,\hat{j}_k}_{k-1}(l_2,l_2)}} > z(\alpha_2)$ **then**
      $K^{\hat{j}^{true}_k}(i,i) = 1$, $\hat{d}^{\hat{j}^{true}_k}_{2,k-1}(l_2) = \hat{d}^{\hat{j}_k}_{2,k-1}(l_2)$;
    else
      $\hat{d}^{\hat{j}^{ture}_k}_{2,k-1}(l_2) = 0$;
    end if
    $l_2 = l_2 + 1$;
  else
    if $\frac{|\hat{d}^{\hat{j}_k}_{1,k}(l_1)|}{\sqrt{P^{d_1,\hat{j}_k}_{k}(l_1,l_1}} > z(\alpha_1)$ **then**
      $K^{\hat{j}^{true}_k}(i,i) = 1$, $\hat{d}^{\hat{j}^{true}_k}_{1,k}(l_1) = \hat{d}^{\hat{j}_k}_{1,k}(l_1)$;
    else
      $\hat{d}^{\hat{j}^{true}_k}_{1,k}(l_1) = 0$;
    end if
    $l_1 = l_1 + 1$;
  end if
end for
Obtain $\hat{j}^{true}_k$ for corresponding $K^{\hat{j}^{true}_k}$;

Return: $\hat{j}^{true}_k$, $\hat{d}^{\hat{j}^{ture}_k}_{1,k}$, $\hat{d}^{\hat{j}^{ture}_k}_{2,k-1}$.

Algorithm 5 NISME with reduced mode set

Input: $\hat{x}^j_{0|0} = \mathbb{E}[x_0]$, $P^{x,j}_0 = P_0$, $\mu^j_0 = \frac{1}{|\mathbb{M}|}$, $\hat{d}^j_{1,0} = (\Sigma^j)^{-1}(y^j_{1,0} - T^j_{1,0}h(\hat{x}^j_{0|0}, u_0, 0, j, 0))$ for $\forall j \in \mathbb{M}$; Choose $0 < \delta \ll \frac{1}{|\mathbb{M}|}$, $0 < \alpha_1 < 1$, $0 < \alpha_2 < 1$, and $\mathcal{T}$;

for $q = 0 : N$ **do**

  $c_1 = 1 + q\mathcal{T}$, $c_2 = (q+1)\mathcal{T}$, $\mathbb{M}^D_{[c_1,c_2]} = \emptyset$;

  for $i \in \mathbb{M}^I$ **do**

    Run Algorithm 3 with ($\mathbb{M}^A$, A^i_k, C^i_k for $k = c_1, c_1+1, \cdots, c_2$, G^j, H^j for $\forall j \in \mathbb{M}^A$) to obtain $\mathbb{M}^{d,i}_{[c_1,c_2]}$;

    $\mathbb{M}^D_{[c_1,c_2]} = \mathbb{M}^D_{[c_1,c_2]} \cup \mathbb{M}^{d,i}_{[c_1,c_2]}$;

  end for

  for $k = c_1 : c_2$ **do**

    Read sensor output y_k, and control input $u(t)$ for $t \in [t_{k-1}, t_k]$;

    for $j \in \mathbb{M}^D_{[c_1,c_2]}$ **do**

      Run the NISE with input $(j, \hat{x}^j_{k-1|k-1}, \hat{d}^j_{1,k-1}, P^{x,j}_{k-1}, P^{d_1,j}_{k-1}, P^{xd_1,j}_{k-1}, y_k, u(t)$ for $t \in [t_{k-1}, t_k])$ to generate output $(\hat{x}^j_{k|k}$, $\hat{d}^j_{1,k}$, $\hat{d}^j_{2,k-1}$, $P^{x,j}_k$, $P^{d_2,j}_{k-1}$, $P^{d_1,j}_k$, $P^{xd_1,j}_k$, $\mathcal{N}^j_k)$;

    end for

▷ ***Mode estimator***

    for $j \in \mathbb{M}^D_{[c_1,c_2]}$ **do**

      $\bar{\mu}^j_k = \max\{\frac{\mathcal{N}^j_k \mu^j_{k-1}}{\sum_{i=1}^{|\mathbb{M}^D_{[c_1,c_2]}|} \mathcal{N}^i_k \mu^i_{k-1}}, \delta\}$;

    end for

    for $j \in \mathbb{M}^D_{[c_1,c_2]}$ **do**

      $\mu^j_k = \frac{\bar{\mu}^j_k}{\sum_{i=1}^{|\mathbb{M}^D_{[c_1,c_2]}|} \bar{\mu}^i_k}$;

    end for

    Set $\hat{j}_k = \text{argmax}_j\, \mu^j_k$;

    Run Algorithm 4 with ($\hat{j}_k$, $\hat{d}^{\hat{j}_k}_{1,k}$, $\hat{d}^{\hat{j}_k}_{2,k-1}$, $P^{d_1,\hat{j}_k}_k$, $P^{d_2,\hat{j}_k}_{k-1}$, α_1, α_2) to obtain ($\hat{j}^{true}_k$, $\hat{d}^{\hat{j}^{ture}_k}_{1,k}$, $\hat{d}^{\hat{j}^{ture}_k}_{2,k-1}$);

    Return:

$$\hat{j}(t) = \hat{j}^{true}_k \text{ for } t \in (t_{k-1}, t_k],\ \hat{x}(t) = \hat{x}^{\hat{j}_k}_{k|k}, t \in (t_{k-1}, t_k],$$

$$\hat{d}_1(t) = \hat{d}^{\hat{j}^{ture}_k}_{1,k} \text{ for } t \in [t_k, t_{k+1}),\ \hat{d}_2(t) = \hat{d}^{\hat{j}^{ture}_k}_{2,k-1} \text{ for } t \in (t_{k-1}, t_k].$$

  end for

end for

2.7 Conclusion

We formulate the attack-resilient estimation of a class of switched nonlinear stochastic systems as the problem of joint estimation of the states, attack vectors and modes. The proposed estimator, the NISME, consists of multiple NISE and a mode estimator. Each NISE is able to generate state and attack estimates for a particular mode and the mode estimator chooses the most likely one. Lastly, the NISME uses the estimates of the selected mode as outputs. We formally analyze the stability of estimation errors in probability for the proposed estimator associated with the true mode under the time-invariant hidden mode. We propose a way to alleviate computational complexity by reducing the number of modes.

2.8 Appendix: Estimator derivation and analysis

Section 2.8.1 derives the NISE algorithm. The proofs of Theorem 2.5.1 and Lemma 2.6.1 are presented in Section 2.8.3. Gauss Markov Theorem will be used in the derivation of the NISE algorithm.

Theorem 2.8.1 *(Gauss Markov Theorem [103]) Estimate* $\hat{x} = (H^*H)^{-1}H^*y$ *is the unbiased linear estimate with smallest variance (among all linear and unbiased estimates) for the model* $y = Hx + v$ *where* v *is a zero-mean random variable with unit variance and* H *has full column rank.*

2.8.1 Derivation of the NISE algorithm

Covariance matrices used in the NISE algorithm are positive (semi) definite as shown in the following lemma.

Lemma 2.8.1 *If* $P_0^x \geq 0$*, then, for all* $k \geq 1$*,* $P_{k|k-1}^x \geq 0$*, and the following matrices induced by the NISE algorithm are positive definite:* $\bar{P}_{k|k-1} \triangleq C_{3,k}P_{k|k-1}^x C_{3,k}^T + R_{3,k}$*,* $\tilde{R}_{1,k} \triangleq C_{1,k}P_k^x C_{1,k}^T + R_{1,k}$ *(if* $rk(\Sigma_k) \neq 0$*), and*

$$\begin{aligned}&\tilde{R}_{2,k}\\ &\triangleq C_{2,k}(I + \epsilon A_{k-1} - \epsilon G_{1,k-1}M_{1,k}C_{1,k-1})P_{k-1}^x(I + \epsilon A_{k-1} - \epsilon G_{1,k-1}M_{1,k}C_{1,k-1})^T C_{2,k}^T\\ &+ \epsilon^2 C_{2,k}G_{1,k-1}M_{1,k}R_{1,k-1}M_{1,k}^T G_{1,k-1}^T C_{2,k}^T + \epsilon^2 C_{2,k}Q_{k-1}C_{2,k}^T + R_{2,k}\end{aligned}$$

(if $rk(\bar{\Sigma}_k) \neq 0$).

PROOF. We first prove that $P^x_{k-1} \geq 0$, and $P^x_{k|k-1} \geq 0$ for all $k \geq 1$ by induction. It holds that $P^x_0 \geq 0$. Since $P^x_0 \geq 0$ and $\bar{Q}_0 \geq 0$, it holds that $P^x_{1|0} \geq 0$ in line 8 of the NISE algorithm. Assume $P^x_{k-1} \geq 0$, and $P^x_{k|k-1} \geq 0$ for some $k \geq 1$. It follows from $R_{3,k} > 0$ that L_k is well-defined. This implies that $P^x_k \geq 0$ in line 11, and $P^x_{k+1|k} \geq 0$ in line 8 of the NISE algorithm. By induction, we conclude that $P^x_{k-1} \geq 0$, and $P^x_{k|k-1} \geq 0$ for all $k \geq 1$.

Pick any $k \geq 1$. Since $R_{1,k}$ (if $rk(\Sigma_k) \neq 0$), $R_{2,k}$ (if $rk(\bar{\Sigma}_k) \neq 0$), and $R_{3,k}$ are positive definite, it holds that $\bar{P}_{k|k-1}$, $\tilde{R}_{1,k}$, and $\tilde{R}_{2,k}$ are positive definite. This completes the proof. ■

Attack vector $d_{1,k-1}$ estimation. Given $\hat{x}_{k-1|k-1}$, attack vector $d_{1,k-1}$ can be estimated by $y_{1,k-1}$ in (2.9):

$$\begin{aligned}\hat{d}_{1,k-1} &= M_{1,k}(y_{1,k-1} - h_1(\hat{x}_{k-1|k-1}, u_{k-1}, 0, t_{k-1})) \\ &= M_{1,k}(C_{1,k-1}\tilde{x}_{k-1|k-1} + H_{1,k}d_{1,k-1} + v_{1,k-1} + \psi_{1,k-1}) \end{aligned} \tag{2.13}$$

where h_1 is linearized. Assuming $\mathbb{E}[\tilde{x}_{k-1|k-1}] = 0$, $\psi_{1,k-1} = 0$, and normalizing the covariance matrix of the right hand side of (2.13), we can choose gain matrix $M_{1,k}$ by the Gauss Markov theorem (Theorem 2.8.1):

$$M_{1,k} = (H^T_{1,k}\tilde{R}^{-1}_{1,k-1}H_{1,k})^{-1}H^T_{1,k}\tilde{R}^{-1}_{1,k-1} = H^{-1}_{1,k}$$

where $\tilde{R}_{1,k-1}$ is nonsingular by Lemma 2.8.1 if $rk(\Sigma_k) \neq 0$. Estimation error $\tilde{d}_{1,k-1}$ is obtained by

$$\tilde{d}_{1,k-1} = -M_{1,k}(C_{1,k-1}\tilde{x}_{k-1|k-1} + v_{1,k-1} + \psi_{1,k-1}) \tag{2.14}$$

where the property $M_{1,k}H_{1,k} = I$ is used.

Attack vector $d_{2,k-1}$ estimation. Attack vector $d_{2,k-1}$ can be estimated by $y_{2,k}$ in (2.9) as follows:

$$\begin{aligned}&\hat{d}_{2,k-1} = M_{2,k}(y_{2,k} - h_2(u_k, 0, t_k) - C_{2,k}(\hat{x}_{k-1|k-1} \\ &+ \epsilon f(\hat{x}_{k-1|k-1}, u_{k-1}, \hat{d}_{1,k-1}, 0, t_{k-1}))) \\ &= M_{2,k}(C_{2,k}(I + \epsilon A_{k-1})\tilde{x}_{k-1|k-1} + C_{2,k}\rho_{k-1} + \epsilon C_{2,k}G_{1,k-1}\tilde{d}_{1,k-1} + \epsilon C_{2,k}G_{2,k-1}d_{2,k-1}\end{aligned}$$

$$+ \epsilon C_{2,k} w_{k-1} + \epsilon C_{2,k} \phi_{k-1} + v_{2,k} + \psi_{2,k})$$

where functions h_2 and f are linearized. Again, assuming that $\phi_{k-1} = \rho_{k-1} = \psi_{2,k} = 0$, $\mathbb{E}[\tilde{x}_{k-1|k-1}] = 0$, and $\mathbb{E}[\tilde{d}_{1,k-1}] = 0$, we choose gain matrix $M_{2,k}$ by the Gauss Markov theorem (Theorem 2.8.1) as follows:

$$M_{2,k} = (\epsilon^2 G_{2,k-1}^T C_{2,k}^T \tilde{R}_{2,k}^{-1} C_{2,k} G_{2,k-1})^{-1} \epsilon G_{2,k-1}^T C_{2,k}^T \tilde{R}_{2,k}^{-1} = (\epsilon C_{2,k} G_{2,k-1})^{-1} = (\epsilon \bar{\Sigma}_k)^{-1}$$

where $\bar{\Sigma}_k$ and $\tilde{R}_{2,k}$ are nonsingular by Lemma 2.8.1. Estimation error of $d_{2,k-1}$ is given by

$$\begin{aligned} \tilde{d}_{2,k-1} = -M_{2,k}(C_{2,k}(I + \epsilon A_{k-1})\tilde{x}_{k-1|k-1} + \epsilon C_{2,k} G_{1,k-1} \tilde{d}_{1,k-1} + \epsilon C_{2,k} w_{k-1} \\ + \epsilon C_{2,k} \phi_{k-1} + v_{2,k} + \psi_{2,k} + C_{2,k} \rho_{k-1}) \end{aligned} \tag{2.15}$$

where the property $\epsilon M_{2,k} C_{2,k} G_{2,k-1} = I$ is used.

State prediction. Generate state prediction $\hat{x}_{k|k-1}$ by simulating system (2.9) for $t \in (t_{k-1}, t_k]$ as follows:

$$\dot{\hat{x}}(t) = f(\hat{x}(t), u(t), \hat{d}_{1,k-1}, 0, t) + G_{2,k-1} \hat{d}_{2,k-1} \tag{2.16}$$

where initial condition is $\hat{x}(t_{k-1}) = \hat{x}_{k-1|k-1}$ and state prediction is $\hat{x}_{k|k-1} = \hat{x}(t_k)$. Or equivalently,

$$\hat{x}_{k|k-1} = \hat{x}_{k-1|k-1} + \epsilon f(\hat{x}_{k-1|k-1}, u_{k-1}, \hat{d}_{1,k-1}, 0, t_{k-1}) + \epsilon G_{2,k-1} \hat{d}_{2,k-1} + \hat{\rho}_{k-1}.$$

Let $\tilde{\rho}_{k-1} \triangleq \rho_{k-1} - \hat{\rho}_{k-1}$. Using (2.10), the state estimation error becomes

$$\begin{aligned} \tilde{x}_{k|k-1} = (I + \epsilon A_{k-1})\tilde{x}_{k-1|k-1} + \epsilon G_{1,k-1} \tilde{d}_{1,k-1} + \epsilon G_{2,k-1} \tilde{d}_{2,k-1} + \epsilon w_{k-1} + \epsilon \phi_{k-1} \\ + \tilde{\rho}_{k-1}. \end{aligned} \tag{2.17}$$

Substitution (2.14) and (2.15) into (2.17) leads to

$$P_{k|k-1}^x = \bar{A}_{k-1} P_{k-1}^x (\bar{A}_{k-1})^T + \bar{Q}_{k-1} \tag{2.18}$$

where we assume $\phi_{k-1} = \psi_{1,k-1} = \psi_{2,k} = \tilde{\rho}_{k-1} = \rho_{k-1} = 0$.

State estimation. Update state prediction $\hat{x}_{k|k-1}$ as

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + L_k(y_{3,k} - h_3(\hat{x}_{k|k-1}, u_k, 0, t_k)). \tag{2.19}$$

By substituting $y_{3,k}$ in (2.9) into (2.19) and linearizing h_3 as (2.8), the state estimation error becomes

$$\tilde{x}_{k|k} = (I - L_k C_{3,k})\tilde{x}_{k|k-1} - L_k v_{3,k} - L_k \psi_{3,k}. \tag{2.20}$$

Its error covariance matrix is

$$P_k^x = (I - L_k C_{3,k}) P_{k|k-1}^x (I - L_k C_{3,k})^T + L_k R_{3,k} L_k^T \tag{2.21}$$

by assuming $\psi_{3,k} = 0$. Minimizing error covariance $\mathrm{tr}(P_k^x)$ with decision variable L_k is an unconstrained optimization problem. We can find the minimizer by taking derivative of $\mathrm{tr}(P_k^x)$ and setting it equal to zero $\frac{\partial \,\mathrm{tr}(P_k^x)}{\partial L_k} = 2((C_{3,k}P_{k|k-1}^x\ (C_{3,k})^T + R_{3,k})L_k^T - C_{3,k}P_{k|k-1}^x) = 0$. The solution is $L_k = P_{k|k-1}^x C_{3,k}^T (C_{3,k}P_{k|k-1}^x (C_{3,k})^T + R_{3,k})^{-1}$ which is well defined as $P_{k|k-1}^x \geq 0$ by Lemma 2.8.1, and $R_{3,k} > 0$. Substitution of (2.18) into (2.21) yields the error covariance update law:

$$\begin{aligned} P_k^x = {} & (I - L_k C_{3,k})\bar{A}_{k-1} P_{k-1}^x \bar{A}_{k-1}^T (I - L_k C_{3,k})^T + (I - L_k C_{3,k})\bar{Q}_{k-1}(I - L_k C_{3,k})^T \\ & + L_k R_{3,k} L_k^T. \end{aligned} \tag{2.22}$$

Substituting (2.14), (2.15), and (2.17) into (2.20) yields the following update law for the state estimation error:

$$\tilde{x}_{k|k} = (I - L_k C_{3,k})(\bar{A}_{k-1}\tilde{x}_{k-1|k-1} + \bar{w}_{k-1} + \bar{\phi}_{k-1} + \bar{\rho}_{k-1}) - L_k(v_{3,k} + \psi_{3,k}) \tag{2.23}$$

where $\bar{A}_{k-1}$ and $\bar{w}_{k-1}$ are defined in (2.11), and

$$\begin{aligned} \bar{\phi}_{k-1} &\triangleq \epsilon(I - \epsilon G_{2,k-1} M_{2,k} C_{2,k})(\phi_{k-1} - G_{1,k-1} M_{1,k} \psi_{1,k-1}) - \epsilon G_{2,k-1} M_{2,k} \psi_{2,k} \\ \bar{\rho}_{k-1} &\triangleq -\epsilon G_{2,k-1} M_{2,k} C_{2,k} \rho_{k-1} + \tilde{\rho}_{k-1}. \end{aligned}$$

The priori probability of the mode. The priori probability of the mode is derived and explained in the following section.

2.8.2 Derivation of the mode estimator

It is natural that the predicted output must be matched with the measured output if the mode j is the true mode. For $\forall j \in \mathbb{M}$, we quantify the discrepancy between the predicted output and the measured output as follows

$$\nu_k^j = y_{3,k}^j - h_3(\hat{x}_{k|k-1}^j, u_k, 0, j, t_k).$$

We approximate the output error ν_k^j as a multivariate Gaussian random variable. Then, the likelihood function is given by

$$\mathcal{N}_k^j \triangleq \mathcal{P}(y_k|j = \text{true}) = \mathcal{N}(\nu_k^j; 0, \bar{P}_{k|k-1}^j) = \frac{\exp(-(\nu_k^j)^T(\bar{P}_{k|k-1}^j)^{-1}\nu_k^j/2)}{(2\pi)^{n^j/2}|\bar{P}_{k|k-1}^j|^{1/2}}$$

where $\bar{P}_{k|k-1}^j = C_{3,k}^j P_{k|k-1}^j (C_{3,k}^j)^T + R_{3,k}^j$ is the error covariance matrix of ν_k^j and $n^j = Rank(\bar{P}_{k|k-1}^j)$. The likelihood function is well-defined since $\bar{P}_{k|k-1}^j > 0$ as shown in Lemma 2.8.1. By the Bayes' theorem, the posterior probabilities are $\mu_k^j \triangleq \mathcal{P}(j = \text{true}|y_k, \cdots, y_0) = \frac{\mathcal{P}(y_k|j=\text{true})\mathcal{P}(j=\text{true}|y_{k-1},\cdots,y_0)}{\sum_{i=1}^{|\mathbb{M}|}\mathcal{P}(y_k|i=\text{true})\mathcal{P}(i=\text{true}|y_{k-1},\cdots,y_0)} = \frac{\mathcal{N}_k^j\mu_{k-1}^j}{\sum_{i=1}^{|\mathbb{M}|}\mathcal{N}_k^i\mu_{k-1}^i}$. However, such update might allow that some μ_k^j converge to zero. To prevent this, we modify the posterior probability update to $\mu_k^j = \frac{\bar{\mu}_k^j}{\sum_{i=1}^{|\mathbb{M}|}\bar{\mu}_k^i}$, where $\bar{\mu}_k^j = \max\{\frac{\mathcal{N}_k^j\mu_{k-1}^j}{\sum_{i=1}^{|\mathbb{M}|}\mathcal{N}_k^i\mu_{k-1}^i}, \delta\}$ and $\delta > 0$ is a pre-selected small constant preventing the vanishment of the mode probabilities. The last step is to generate the state, attack vector, and mode estimates of the mode having the maximum posteriori probability.

2.8.3 Stability analysis of the NISE algorithm

Without loss of generality, consider constants $\bar{c}_1 = \bar{g}_1 = \bar{m}_1 = 0$, if $rk(\Sigma_k) = 0$ in Assumption 2.5.2. Likewise, if $rk(\bar{\Sigma}_k) = 0$, consider constants $\bar{c}_2 = \underline{g}_2 = \bar{g}_2 = \underline{m}_2 = \bar{m}_2 = \underline{r}_2 = 0$ in this section.

Proof of Theorem 2.5.1. We choose the Lyapunov function candidate $V_k(\tilde{x}_{k|k}) \triangleq \tilde{x}_{k|k}^T(P_k^x)^{-1}\tilde{x}_{k|k}$ where $0 < \underline{p}I \leq P_k^x \leq \bar{p}I$. Substituting (2.23) into $V_k(\tilde{x}_{k|k})$, we have

$$\begin{aligned} V_k(\tilde{x}_{k|k}) &= \tilde{x}_{k-1|k-1}^T\bar{A}_{k-1}^T(I - L_kC_{3,k})^T(P_k^x)^{-1}(I - L_kC_{3,k})\bar{A}_{k-1}\tilde{x}_{k-1|k-1} \\ &+ \bar{v}_k^T(P_k^x)^{-1}(2(I - L_kC_{3,k})\bar{A}_{k-1}\tilde{x}_{k-1|k-1} + \bar{v}_k + 2\bar{\psi}_k) \end{aligned}$$

$$+ \bar{\psi}_k^T (P_k^x)^{-1} (2(I - L_k C_{3,k}) \bar{A}_{k-1} \tilde{x}_{k-1|k-1} + \bar{\psi}_k) \tag{2.24}$$

where $\bar{\psi}_k \triangleq (I - L_k C_{3,k})(\bar{\phi}_{k-1} + \bar{\rho}_{k-1}) - L_k \psi_{3,k}$ and $\bar{v}_k \triangleq (I - L_k C_{3,k}) \bar{w}_{k-1} - L_k v_{3,k}$. Assumption 2.5.2 implies that $\|\bar{A}_k\| \leq \bar{a} \triangleq (1 + \bar{g}_2 \bar{c}_2 \bar{m}_2)(1 + \epsilon \bar{a}' + \epsilon \bar{g}_1 \bar{c}_1 \bar{m}_1)$, and $\underline{q} I \leq \bar{Q}_k \leq \bar{q} I$ where $\underline{q} \triangleq \max\{\underline{g}_2^2 \underline{m}_2^2 \underline{r}_2, \epsilon^2 \underline{q}'\}$ and $\bar{q} \triangleq \epsilon^2 (1 + \bar{g}_2 \bar{c}_2 \bar{m}_2)^2 (\bar{q}' + \bar{g}_1 \bar{m}_1 \bar{r}_1)^2 + \bar{g}_2^2 \bar{m}_2^2 \bar{r}_2$.

Claims 1-3 derive bounds for the first to third terms in (2.24), respectively.

Claim 1: There exists constant $\alpha' \triangleq (1 + \frac{\underline{q}}{\bar{a}^2 \bar{p}})^{-1} \in (0, 1)$ such that $\bar{A}_{k-1}^T (I - L_k C_{3,k})^T (P_k^x)^{-1} (I - L_k C_{3,k})\ \bar{A}_{k-1} < \alpha' (P_{k-1}^x)^{-1}$.

PROOF. It follows from Assumptions 2.5.2 and 2.5.4 that

$$\bar{Q}_{k-1} \geq \underline{q} I,\ \frac{P_{k-1}^x}{\bar{p}} \leq I. \tag{2.25}$$

Since $\bar{A}_{k-1} \bar{A}_{k-1}^T$ is symmetric, it holds that

$$\bar{A}_{k-1} \bar{A}_{k-1}^T \leq \lambda_{\max}(\bar{A}_{k-1} \bar{A}_{k-1}^T) I = \bar{a}^2 I. \tag{2.26}$$

Thus, by (2.25) and (2.26), we have

$$\bar{Q}_{k-1} \geq \frac{\underline{q}}{\bar{a}^2} \bar{A}_{k-1} \bar{A}_{k-1}^T \geq \frac{\underline{q}}{\bar{a}^2 \bar{p}} \bar{A}_{k-1} P_{k-1}^x \bar{A}_{k-1}^T. \tag{2.27}$$

Substitution of (2.27) to (2.22) yields

$$P_k^x \geq (I - L_k C_{3,k}) \bar{A}_{k-1} P_{k-1}^x \bar{A}_{k-1}^T (I - L_k C_{3,k})^T (1 + \frac{\underline{q}}{\bar{a}^2 \bar{p}}) + L_k R_{3,k} L_k^T.$$

This implies $P_k^x - (I - L_k C_{3,k}) \bar{A}_{k-1} P_{k-1}^x \bar{A}_{k-1}^T (I - L_k C_{3,k})^T \geq 0$. Since $P_{k-1}^x > 0$, and then

$$\begin{aligned}
&(1 + \frac{\underline{q}}{\bar{a}^2 \bar{p}}) P_{k-1}^x + (1 + \frac{\underline{q}}{\bar{a}^2 \bar{p}})^2 P_{k-1}^x \bar{A}_{k-1}^T (I - L_k C_{3,k})^T \\
&\times (P_k^x - (I - L_k C_{3,k}) \bar{A}_{k-1} P_{k-1}^x \bar{A}_{k-1}^T (I - L_k C_{3,k})^T)(I - L_k C_{3,k}) \bar{A}_{k-1} P_{k-1}^x \\
&\geq (1 + \frac{\underline{q}}{\bar{a}^2 \bar{p}}) P_{k-1}^x > 0.
\end{aligned}$$

By the matrix inversion lemma [101], it follows that

$$((1+\frac{\underline{q}}{\bar{a}^2\bar{p}})^{-1}(P_{k-1}^x)^{-1} - \bar{A}_{k-1}^T(I-L_kC_{3,k})^T(P_k^x)^{-1}(I-L_kC_{3,k})\bar{A}_{k-1})^{-1} > 0.$$

It establishes the statement with $\alpha' = (1+\frac{\underline{q}}{\bar{a}^2\bar{p}})^{-1}$. ■

Claim 2: There exists constant $\epsilon_1 > 0$ such that

$$\mathbb{E}[\bar{v}_k^T(P_k^x)^{-1}(2(I-L_kC_{3,k})\bar{A}_{k-1}\tilde{x}_{k-1|k-1} + \bar{v}_k + 2\bar{\psi}_k)] \leq \epsilon_1.$$

PROOF. Noises $w_{k-1}, v_{1,k-1}, v_{2,k}$, and $v_{3,k}$ are uncorrelated and thus we have

$$\begin{aligned}
&\mathbb{E}[\bar{v}_k^T(P_k^x)^{-1}(2(I-L_kC_{3,k})\bar{A}_{k-1}\tilde{x}_{k-1|k-1} + \bar{v}_k + 2\bar{\psi}_k)] \\
&= \mathbb{E}[\bar{v}_k^T(P_k^x)^{-1}\bar{v}_k] \leq \epsilon^2\underline{p}(1+\bar{l}\bar{c}_3)^2((1+\bar{g}_2\bar{m}_2\bar{c}_2)^2(\bar{q}rk(Q_{k-1}) + \bar{g}_1^2\bar{m}_1^2\bar{r}_1rk(R_{1,k-1})) \\
&+ \bar{g}_2^2\bar{m}_2^2\bar{r}_2rk(R_{2,k})) + \underline{p}\bar{l}^2\bar{r}_3rk(R_{3,k}) \triangleq \epsilon_1
\end{aligned}$$

where we apply $\|v_{1,k}\|^2 = \mathrm{tr}(v_{1,k}^Tv_{1,k}) = \mathrm{tr}(v_{1,k}v_{1,k}^T) \leq \bar{r}_1rk(R_{1,k})$ and the similar relations for $\|w_{k-1}\|^2$, $\|v_{2,k}\|^2$, and $\|v_{3,k}\|^2$. ■

Claim 3: There exist constants $\delta, \delta_\rho, \lambda, \epsilon_2 > 0$ such that, for $\forall\|\tilde{x}_{k-1|k-1}\| \leq \delta$ and $\epsilon \leq \delta_\rho$, the following holds:

$$\bar{\psi}_k^T(P_k^x)^{-1}(2(I-L_kC_{3,k})\bar{A}_{k-1}\tilde{x}_{k-1|k-1} + \bar{\psi}_k) \leq \lambda\|\tilde{x}_{k-1|k-1}\|^3 + \epsilon_2.$$

PROOF. By Assumptions 2.5.2, 2.5.3 and 2.5.5, it holds that

$$\begin{aligned}
\|\bar{\psi}_k\| &= \|(I-L_kC_{3,k})(\bar{\phi}_{k-1} + \bar{\rho}_{k-1}) - L_k\psi_{3,k}\| \\
&\leq (1+\bar{l}\bar{c}_3)(\epsilon((1+\bar{g}_2\bar{c}_2\bar{m}_2)(\epsilon_\phi + \epsilon_{\psi_1}\bar{g}_1\bar{m}_1) + \bar{g}_2\bar{m}_2\epsilon_{\psi_2})\|\tilde{x}_{k-1|k-1}\|^2 \\
&+ \epsilon^2\epsilon_\rho(\bar{g}_2\bar{m}_2\bar{c}_2 + 2)) + \bar{l}\epsilon_{\psi_3}\|\tilde{x}_{k-1|k-1}\|^2 \\
&\triangleq \lambda'\|\tilde{x}_{k-1|k-1}\|^2 + \epsilon_2'
\end{aligned}$$

for all $\|\tilde{x}_{k-1|k-1}\| \leq \delta$, and $\epsilon \leq \delta_\rho$. Hence,

$$\begin{aligned}
&\bar{\psi}_k^T(P_k^x)^{-1}(2(I-L_kC_{3,k})\bar{A}_{k-1}\tilde{x}_{k-1|k-1} + \bar{\psi}_k) \\
&\leq (\lambda'\|\tilde{x}_{k-1|k-1}\|^2 + \epsilon_2')(P_k^x)^{-1}(2(a+\bar{l}\bar{c}_3)\bar{a}\|\tilde{x}_{k-1|k-1}\| + \lambda'\|\tilde{x}_{k-1|k-1}\|^2 + \epsilon_2') \\
&\leq 2\lambda'\underline{p}(1+\bar{l}\bar{c}_3)\bar{a}\|\tilde{x}_{k-1|k-1}\|^3 + \lambda'^2\underline{p}\delta\|\tilde{x}_{k-1|k-1}\|^3 + \epsilon_2'\underline{p}(2(1+\bar{l}\bar{c}_3)\bar{a}\delta + 2\lambda'\delta^2 + \epsilon_2')
\end{aligned}$$

$\triangleq \lambda\|\tilde{x}_{k-1|k-1}\|^3 + \epsilon_2$

where $\|\tilde{x}_{k-1|k-1}\| \leq \delta$ is applied in the last inequality. ■

By Claims 1-3, recursion (2.24) becomes

$$\mathbb{E}[V_k(\tilde{x}_{k|k})] \leq \alpha'\mathbb{E}[\tilde{x}_{k-1|k-1}^T(P_{k-1}^x)^{-1}\tilde{x}_{k-1|k-1}] + \lambda\mathbb{E}[\|\tilde{x}_{k-1|k-1}\|^3] + (\epsilon_1 + \epsilon_2)$$

for $\forall\|\tilde{x}_{k-1|k-1}\| \leq \delta$ and $\epsilon \leq \delta_\rho$. Notice that λ tends to zero as $\epsilon_\phi, \epsilon_{\psi_1}, \epsilon_{\psi_2}, \epsilon_{\psi_3}$ tend to zero. Thus, there exists a sufficiently small tuple $(\epsilon_\phi, \epsilon_{\psi_1}, \epsilon_{\psi_3})$ such that $\lambda\delta < \alpha'\bar{p}^{-1}$. Then,

$$\mathbb{E}[V_k(\tilde{x}_{k|k})] \leq \alpha\mathbb{E}[V_{k-1}(\tilde{x}_{k-1|k-1})] + c \tag{2.28}$$

for $\|\tilde{x}_{k-1|k-1}\| \leq \delta$ where $0 < \alpha < 1$ and $c \triangleq \epsilon_1 + \epsilon_2$. Remind that ϵ_1 tends to zero as $\bar{q}'$, $\bar{r}_1$, $\bar{r}_2$ and $\bar{r}_3$ tend to zero, and constant ϵ_2 tends to zero as ϵ tends to zero. For any constant $c' > 0$, there exists a sufficiently small tuple $(\bar{q}', \bar{r}_1, \bar{r}_2, \bar{r}_3, \bar{\epsilon})$ such that $c < c'$ holds for all $\epsilon \leq \bar{\epsilon}$. The following claim shows the PESp-like property for state estimation error $\tilde{x}_{k|k}$.

Claim 4: For any $\gamma \in (0, 1)$, there exist α_x, b_x, c_x, $\underline{\delta} > 0$ and tuple $(\bar{q}', \bar{r}_1, \bar{r}_2, \bar{r}_3, \bar{\epsilon})$ such that if $Q_k \leq \bar{q}'I$, $R_{1,k} \leq \bar{r}_1 I$, $R_{2,k} \leq \bar{r}_2 I$, $R_{3,k} \leq \bar{r}_3 I$, and $\epsilon \leq \bar{\epsilon}$, then the following properties hold, for all $\|\tilde{x}_{0|0}\| \leq \underline{\delta}$ and $k \geq 0$:

$$P(\|\tilde{x}_{k|k}\| < \alpha_x e^{-b_x k}\|\tilde{x}_{0|0}\| + c_x) \geq 1 - \gamma.$$

PROOF. Consider any $\gamma \in (0, 1)$ and $\gamma_1 < \gamma$. Then, there exists sufficiently small constant $\underline{\delta} < \delta$ and tuple $(\bar{q}', \bar{r}_1, \bar{r}_2, \bar{r}_3, \bar{\epsilon})$ such that $\bar{p}\underline{\delta}^2 + \frac{c}{1-\alpha} \leq \gamma_1\underline{p}\underline{\delta}^2$ holds. Since $V_0(\tilde{x}_{0|0}) \leq \bar{p}\|\tilde{x}_{0|0}\|^2$, we have, for all $\|\tilde{x}_{0|0}\| \leq \underline{\delta}$,

$$V_0(\tilde{x}_{0|0}) + \frac{c}{1-\alpha} \leq \bar{p}\|\tilde{x}_{0|0}\|^2 + \frac{c}{1-\alpha} \leq \gamma_1\underline{p}\underline{\delta}^2. \tag{2.29}$$

We choose any $\|\tilde{x}_{0|0}\| \leq \underline{\delta}$. Define the first exit time $\mu \triangleq \inf\{t_k > 0 | \|\tilde{x}_{k|k}\| > \delta\}$, and $\mu \wedge k \triangleq \min\{\mu, k\}$ for any $k > 0$. We have

$$\underline{p}\underline{\delta}^2 P(\mu \leq k) \leq \mathbb{E}[V_\mu(\tilde{x}_{\mu|\mu})I_{[\mu\leq k]}] \leq \mathbb{E}[V_{\mu\wedge k}(\tilde{x}_{\mu\wedge k|\mu\wedge k})]$$

$$\leq \alpha^{\mu\wedge k}V_0(\tilde{x}_{0|0}) + c\sum_{i=0}^{\mu\wedge k-1}\alpha^i \tag{2.30}$$

where indicator function $I_{[\mu\leq k]}$ satisfies $I_{[\mu\leq k]} = 1$ if $\mu \leq k$, otherwise $I_{[\mu\leq k]} = 0$. The first inequality holds because $\underline{p}\delta^2 < \underline{p}\|\tilde{x}_{\mu|\mu}\|^2 \leq V_\mu(\tilde{x}_{\mu|\mu})$. The third inequality can be obtained by recursively applying (2.28). It follows from (2.29) and (2.30) that

$$P(\mu \leq k) \leq \gamma_1 \frac{\alpha^{\mu\wedge k}V_0(\tilde{x}_{0|0}) + c\sum_{i=0}^{\mu\wedge k-1}\alpha^i}{V_0(\tilde{x}_{0|0}) + \frac{c}{1-\alpha}} \leq \gamma_1. \tag{2.31}$$

Letting $k \to \infty$, we also have $P(\|\tilde{x}_{k|k}\| \leq \delta) \geq 1 - \gamma_1$. Again consider any k with $\mu \wedge k$, and $\gamma_2 \triangleq \gamma - \gamma_1$. Markov's inequality (p.455 [102]) derives

$$\begin{aligned} &P(V_{\mu\wedge k}(\tilde{x}_{\mu\wedge k|\mu\wedge k}) \geq \frac{\alpha^{\mu\wedge k}V_0(\tilde{x}_{0|0}) + \frac{c}{1-\alpha}}{\gamma_2}) \leq \gamma_2 \frac{\mathbb{E}[V_{\mu\wedge k}(\tilde{x}_{\mu\wedge k|\mu\wedge k})]}{\alpha^{\mu\wedge k}V_0(\tilde{x}_{0|0}) + \frac{c}{1-\alpha}} \\ &\leq \gamma_2 \frac{\alpha^{\mu\wedge k}V_0(\tilde{x}_{0|0}) + c\sum_{i=0}^{\mu\wedge k-1}\alpha^i}{\alpha^{\mu\wedge k}V_0(\tilde{x}_{0|0}) + \frac{c}{1-\alpha}} \leq \gamma_2. \end{aligned}$$

Equivalently, $P(V_{\mu\wedge k}(\tilde{x}_{\mu\wedge k|\mu\wedge k}) < \frac{\alpha^{\mu\wedge k}V_0(\tilde{x}_{0|0}) + \frac{c}{1-\alpha}}{\gamma_2}) \geq 1 - \gamma_2$. This implies that, by Minkowski inequality,

$$P(\|\tilde{x}_{\mu\wedge k|\mu\wedge k}\| < \beta_x(\|\tilde{x}_{0|0}\|, \mu\wedge k) + c_x) \geq 1 - \gamma_2 \tag{2.32}$$

where $\beta_x(\|\tilde{x}_{0|0}\|, \mu\wedge k) \triangleq \sqrt{\frac{\bar{p}}{\gamma_2\underline{p}}}\alpha^{(\mu\wedge k)/2}\|\tilde{x}_{0|0}\|$ and $c_x \triangleq \sqrt{\frac{c}{(1-\alpha)(\underline{p}\gamma_2)}}$). By (2.31) and (2.32), we can obtain

$$\begin{aligned} &P(\|\tilde{x}_{k|k}\| < \beta_x(\|\tilde{x}_{0|0}\|, k) + c_x) \\ &= P(\|\tilde{x}_{k|k}\| < \beta_x(\|\tilde{x}_{0|0}\|, k) + c_x | \mu > k)P(\mu > k) \\ &\quad + P(\|\tilde{x}_{k|k}\| < \beta_x(\|\tilde{x}_{0|0}\|, k) + c_x | \mu \leq k)P(\mu \leq k) \\ &\geq P(\|\tilde{x}_{\mu\wedge k|\mu\wedge k}\| < \beta_x(\|\tilde{x}_{0|0}\|, \mu\wedge k) + c_x | \mu > k)P(\mu > k) \\ &= P(\|\tilde{x}_{\mu\wedge k|\mu\wedge k}\| < \beta_x(\|\tilde{x}_{0|0}\|, \mu\wedge k) + c_x) \\ &\quad - P(\|\tilde{x}_{\mu\wedge k|\mu\wedge k}\| < \beta_x(\|\tilde{x}_{0|0}\|, \mu\wedge k) + c_x | \mu \leq k)P(\mu \leq k) \\ &\geq 1 - \gamma_2 - \gamma_1 = 1 - \gamma. \end{aligned} \tag{2.33}$$

■

We are now in a position to prove PESp-like property for $\tilde{d}_1(t)$ and $\tilde{d}_2(t)$. The NISE algorithm treats attack vectors as constants $\hat{d}_1(t) = \hat{d}_{1,k}$ and $\hat{d}_2(t) = \hat{d}_{2,k}$ for $t \in [t_k, t_{k+1})$. Let us define constant approximation errors $d_{1,e}(t) \triangleq d_1(t) - \hat{d}_{1,k} - \tilde{d}_{1,k}$ and $d_{2,e}(t) \triangleq d_2(t) - \hat{d}_{2,k} - \tilde{d}_{2,k}$ for $t = [t_k, t_{k+1})$.

Claim 5: The constant approximation errors are bounded: $\|d_{1,e}(t)\| \leq \epsilon\bar{d}$, $\|d_{2,e}(t)\| \leq \epsilon\bar{d}$.

PROOF. Notice that $d_{1,e}(t)$ and $d_{2,e}(t)$ satisfy $\|d_{1,e}(t_k)\| = \|d_{2,e}(t_k)\| = 0$ for $\forall k$ because new estimates are obtained at the sampling instants. Since attack vector $d(t)$ is continuous, the error bound of $d_{1,e}(t)$ is given by

$$\|d_{1,e}(t_1)\| \leq (t_1 - t_k) \sup_{t \in [t_k, t_{k+1})} \|\frac{d(t) - d(t_k)}{t - t_k}\| \leq \epsilon\bar{d}$$

where Assumption 2.5.1 and $t_1 - t_k \leq \epsilon$ are applied. The proof for $d_{2,e}(t)$ is analogous to that for $d_{1,e}(t)$. ■

Without loss of generality, consider γ, γ_1, and $\tilde{x}_{0|0}$ used in Claim 4. Consider mean square error of $\hat{d}_1(t)$ for $t \in [t_k, t_{k+1})$, with (2.14) and $\|d_{1,e}(t)\| \leq \epsilon\bar{d}$ by Claim 5:

$$\begin{aligned}
&\mathbb{E}[\|\tilde{d}_1(\mu \wedge t)\|^2] = \mathbb{E}[\|\tilde{d}_{1,\mu\wedge k} + d_{1,e}(\mu \wedge t)\|^2] \\
&\leq \bar{m}_1^2(\mathbb{E}[\bar{c}_1^2\|\tilde{x}_{\mu\wedge k|\mu\wedge k}\|^2 + \bar{c}_1\epsilon\bar{d}\|\tilde{x}_{\mu\wedge k|\mu\wedge k}\| + \|v_{1,\mu\wedge k}\|^2 + \|\psi_{1,\mu\wedge k}\|^2 \\
&+ 2\bar{c}_1\|\tilde{x}_{\mu\wedge k|\mu\wedge k}\|\|\psi_{1,\mu\wedge k}\| + \epsilon\bar{d}\|\psi_{1,\mu\wedge k}\|]) + \epsilon^2\bar{d}^2
\end{aligned}$$

where Cauchy-Schwarz inequality is applied. Since $\|\tilde{x}_{\mu\wedge k|\mu\wedge k}\| \leq \delta$, it follows that

$$\begin{aligned}
&\mathbb{E}[\|\tilde{d}_1(\mu \wedge t)\|^2] \leq \bar{m}_1^2(\bar{c}_1^2 + 2\epsilon_{\psi_1}\bar{c}_1\delta + \epsilon_{\psi_1}^2(\delta^2 + \epsilon\bar{d}))\mathbb{E}[\|\tilde{x}_{\mu\wedge k|\mu\wedge k}\|^2] \\
&+ \bar{m}_1^2(\bar{c}_1\epsilon\bar{d}\delta + \bar{r}_1 rk(R_{1,\mu\wedge k})) + \epsilon^2\bar{d}^2
\end{aligned}$$

where $\|v_{1,\mu\wedge k}\|^2 \leq \bar{r}_1 rk(R_{1,\mu\wedge k})$, and $\|\psi_{1,\mu\wedge k}\| \leq \epsilon_{\psi_1}\|\tilde{x}_{\mu\wedge k|\mu\wedge k}\|^2$ are applied. Applying $\epsilon \leq c$ and $\mathbb{E}[\|\tilde{x}_{\mu\wedge k|\mu\wedge k}\|^2] \leq \frac{\alpha^{\mu\wedge k}}{\underline{p}} V_0(\tilde{x}_{0|0}) + \frac{c}{\underline{p}} \sum_{i=0}^{\mu\wedge k-1} \alpha^i$ obtained in (2.30), it follows that

$$\mathbb{E}[\|\tilde{d}_1(\mu \wedge t)\|^2] \leq \beta_1(\|\tilde{x}_{0|0}\|^2, \mu \wedge k) + c_1$$

where $\beta_1(\|\tilde{x}_{0|0}\|^2, \mu \wedge k) \triangleq \bar{m}_1^2(\bar{c}_1^2 + 2\epsilon_{\psi_1}\bar{c}_1\delta + \epsilon_{\psi_1}^2(\delta^2 + c\bar{d}))\frac{\alpha^{\mu\wedge k}}{\underline{p}}\bar{p}\|\tilde{x}_{0|0}\|^2$ and $c_1 \triangleq \bar{m}_1^2(\bar{c}_1^2+2\epsilon_{\psi_1}\bar{c}_1\delta+\epsilon_{\psi_1}^2(\delta^2+c\bar{d}))\frac{c}{\underline{p}(1-\alpha)}+\bar{m}_1^2(\bar{c}_1 c\bar{d}\delta+\bar{r}_1 rk(R_{1,\mu\wedge k}))+\epsilon^2\bar{d}^2$. By Markov's inequality, for $t \in [t_k, t_{k+1})$, we have

$$P(\|\tilde{d}_1(\mu \wedge t)\|^2 \geq \frac{\beta_1(\|\tilde{x}_{0|0}\|^2, \mu \wedge k) + c_1}{\gamma_3}) \leq \gamma_3$$

for any $\gamma_3 \in (0, 1)$. Analogous to (2.33), we have

$$\begin{aligned}
&P(\|\tilde{d}_1(t)\|^2 < \frac{\beta_1(\|\tilde{x}_{0|0}\|^2, k) + c_1}{\gamma_3}) \\
&\geq P(\|\tilde{d}_1(\mu \wedge t)\|^2 < \frac{\beta_1(\|\tilde{x}_{0|0}\|^2, \mu \wedge k) + c_1}{\gamma_3} | \mu > k) P(\mu > k) \\
&= P(\|\tilde{d}_1(\mu \wedge t)\|^2 < \frac{\beta_1(\|\tilde{x}_{0|0}\|^2, \mu \wedge k) + c_1}{\gamma_3}) \\
&\quad - P(\|\tilde{d}_1(\mu \wedge t)\|^2 < \frac{\beta_1(\|\tilde{x}_{0|0}\|^2, \mu \wedge k) + c_1}{\gamma_3} | \mu \leq k) P(\mu \leq k) \\
&\geq 1 - \gamma_3 - \gamma_1 \geq \gamma
\end{aligned}$$

for $t \in [t_k, t_{k+1})$ and for some $\gamma \in (0, 1)$. By applying Minkowski inequality to $\frac{\beta_1(\|\tilde{x}_{0|0}\|^2, k)+c_1}{\gamma_3}$, we show PESp-like property for $\tilde{d}_1(t)$. The proof of PESp-like property for $\tilde{d}_2(t)$ is similar to that for $\tilde{d}_1(t)$. We omit its details. ■

Proof of Lemma 2.6.1. To prove the statement, we first formally establish the equivalence of the NISE algorithm and the EKF by expressing the attack vector estimates as functions of state estimates. Due to such connection, we apply existing results on the analysis of the EKF [93] to the NISE algorithm to prove the rest of the part. By expressing $\tilde{d}_{1,k-1}$ and $\tilde{d}_{2,k-1}$ as functions of $\tilde{x}_{k-1|k-1}$, we find the update law of state estimation (2.23) and its error covariance (2.22). They are identical to those of the EKF problem for the following continuous-discrete system:

$$\begin{aligned}
\dot{x}(t) &= \bar{f}(x(t), u(t), \bar{w}'(t), t) \\
y_{3,k} &= \bar{h}(x_k, u_k, v'_k, t_k)
\end{aligned} \tag{2.34}$$

where its linearized system is given by

$$x_{k+1} = x_k + \epsilon\bar{f}(x_k, u_k, \bar{w}'_k, t_k) + \bar{\rho}_k$$

$$\simeq x_k + \epsilon(\bar{A}_k x_k + \bar{B}_k u_k) + \bar{\rho}_k + \bar{w}_k$$
$$y_{3,k} \simeq C_{3,k} x_k + D_{3,k} u_k + v_{3,k}$$

as shown in Claim 6; i.e., the two problems are equivalent to each other.

Claim 6: Under Assumption 2.4.1, state estimation error update law and its error covariance update law of the continuous-discrete EKF problem (2.34) is identical to (2.23) and (2.22) of system (2.3). Moreover, the optimal estimate gain matrices L_k are identical.

PROOF. Consider (2.34). State prediction $\hat{x}_{k|k-1}$ can be obtained by setting $\hat{x}_{k|k-1} = x(t_k)$ where

$$\dot{\hat{x}}(t) = \bar{f}(\hat{x}(t), u(t), 0, t)$$

for $t \in (t_{k-1}, t_k]$ with initial condition $\hat{x}(t_{k-1}) = \hat{x}_{k-1|k-1}$. Or equivalently, $\hat{x}_{k|k-1} = \hat{x}_{k-1|k-1} + \epsilon \bar{f}(\hat{x}_{k-1|k-1}, u_{k-1}, 0, t_{k-1}) + \bar{\rho}_{k-1}$. Its error dynamic is given by, from the linearization of $\bar{f}$,

$$\tilde{x}_{k|k-1} = (I + \epsilon \bar{A}_{k-1})\tilde{x}_{k-1|k-1} + \bar{w}_{k-1} + \bar{\phi}_{k-1}.$$

State estimate is

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + L_k(y_{3,k} - \bar{h}(x_{k|k-1}, u_k, 0, t_k))$$

with its error dynamic

$$\begin{aligned}
\tilde{x}_{k|k} &= (I - L_k C_{3,k})\tilde{x}_{k|k-1} - L_k v_{3,k} - L_k \psi_{3,k} \\
&= \bar{A}_{k-1}\tilde{x}_{k-1|k-1} + \bar{w}_{k-1} + \bar{\phi}_{k-1} + \bar{\rho}_{k-1} \\
&- L_k(C_{3,k}(\bar{A}\tilde{x}_{k-1|k-1} + \bar{w}_{k-1} + \bar{\phi}_{k-1} + \bar{\rho}_{k-1}) + v_{3,k} + \psi_{3,k}).
\end{aligned}$$

which is identical to (2.23). We can find error covariance update (2.22) by the same argument for the NISE algorithm, ignoring unknown terms $\bar{\rho}_{k-1}$, $\bar{\phi}_{k-1}$ and $\psi_{3,k}$. Moreover, it should be emphasized that finding the optimal gain matrices L_k for the EKF is the same unconstrained optimization problem of the NISE algorithm. ■

We are now in a position to verify that all the assumptions on the EKF in [93]

are satisfied. As shown in Claim 6, the NISE algorithm satisfies that process noise $\bar{w}'(t)$ and measurement noise v'_k are uncorrelated in the equivalent EKF problem (2.34). Assumption 2.5.2 implies that $\|\bar{A}_k\| \leq \bar{a}$, $\underline{q}I \leq \bar{Q}_k$ and uniform observability of the pair $(C_{3,k}, \bar{A}_k)$. The above conditions with Assumptions 2.5.2, and 2.5.3 suffice the conditions of Theorem 4.5 in [93]. Because of the equivalence of the EKF and the NISE algorithm, Theorem 4.5 of [93] implies that $\underline{p}I \leq P_k^x = \mathbb{E}[\tilde{x}_k \tilde{x}_k^T] \leq \bar{p}I$ of the NISE algorithm for some positive constants $\underline{p}$ and $\bar{p}$.

Now we proceed to prove the boundedness of $P_k^{d_1}$ and $P_{k-1}^{d_2}$. Let us consider the following instrumental claim.

Claim 7: If $0 \leq \underline{p}I \leq P \leq \bar{p}I$ and $\|F\| \leq \bar{F}$, then $0 \leq F^T P_k^x F \leq \bar{p}\bar{F}^2 I$.

PROOF. Consider any v and choose $v' = Fv$. Then,

$$v'^T(P_k^x - \bar{p}I)v' \leq 0,\ v'^T(P_k^x - \underline{p}I)v' \geq 0.$$

Hence,

$$v^T(F^T P_k^x F - \bar{p}F^T F)v \leq 0, v^T(F^T P_k^x F - \underline{p}F^T F)v \geq 0.$$

Since $F^T F$ is symmetric, we have

$$v^T F^T P_k^x F v \leq \bar{p}\lambda_{\max}(F^T F)\|v\|^2,$$
$$v^T F^T P_k^x F v \geq \underline{p}\lambda_{\min}(F^T F)\|v\|^2.$$

Notice that $\lambda_{\max}(F^T F) = \|F\|^2 \leq \bar{F}^2$ and $\lambda_{\min}(F^T F) \geq 0$. Thus $0 \leq F^T P_k^x F \leq \bar{p}\bar{F}^2 I$. ■

Substitute $P_{k-1}^{xd_1}$ into $P_{k-1}^{d_2}$ and apply Claim 7 with $0 \leq R_{1,k-1} \leq \bar{r}_1 I$, $0 \leq R_{2,k} \leq \bar{r}_2 I$, $0 \leq Q_{k-1} \leq \bar{q}' I$, and $\underline{p}I \leq P_k^x \leq \bar{p}I$. Then, we have $0 \leq P_{k-1}^{d_2} \leq \bar{p}^{d_2} I$ for some $\bar{p}^{d_2} \geq 0$. Likewise, by Claim 7, there exists non-negative constant $\bar{p}^{d_1}$ such that $0 \leq P_k^{d_1} \leq \bar{p}^{d_1} I$. The above arguments hold for all $k \geq 0$. ■

Chapter 3
Cyber-physical security: Attack-resilient estimation (Applications)

We in this chapter present applications of the NISME designed in Chapter 2 on power systems, mobile robots, and vehicle networks. In particular, we extend the intrusion detection systems into distributed settings in Section 3.3 in which a fleet of vehicles collaborates each other to accomplish attack detection.

3.1 Power systems

In Section 2.5, a set of properties is shown for the true mode under the time-invariant hidden mode. However, we do not have formal guarantees for the cases of time-varying modes. Moreover, the effectiveness of a set of reduced modes discussed in Section 2.6.3 remains unclear. In this section, we will use the IEEE 68-bus test system to empirically illustrate them. The NISME is applied to the IEEE 68-bus test system shown in Figure 3.1. In the network, there are 16 generator buses ($|\mathcal{G}| = 16$), and 52 load buses ($|\mathcal{L}| = 52$). Each local bus is described by (2.1) (as [104]), and (2.2) with $\epsilon = 0.01s$. It is assumed that noises $w(t)$ and v_k are zero mean Gaussian with covariance matrices $Q_i(t) = 0.01^2 I$, and $R_{i,k} = 0.01^4 I$. The parameters are adopted from page 598 in [105]: $D_i = 1$, and $t_{ij} = 1.5$ for $\forall i \in \mathcal{V}$. Angular momentums are $m_i = 10s$ for $i \in \mathcal{G}$ and a larger value $m_i = 100s$ for load buses $i \in \mathcal{L}$. Backstepping inspired stabilizing distributed controllers [106] are applied to the power system. We choose $\delta = 3.3\%$ as a lower bound of probabilities. The attacker could launch 3 sensor attacks, 3 actuator attacks, and 2 switching attacks described in Figure 3.1.

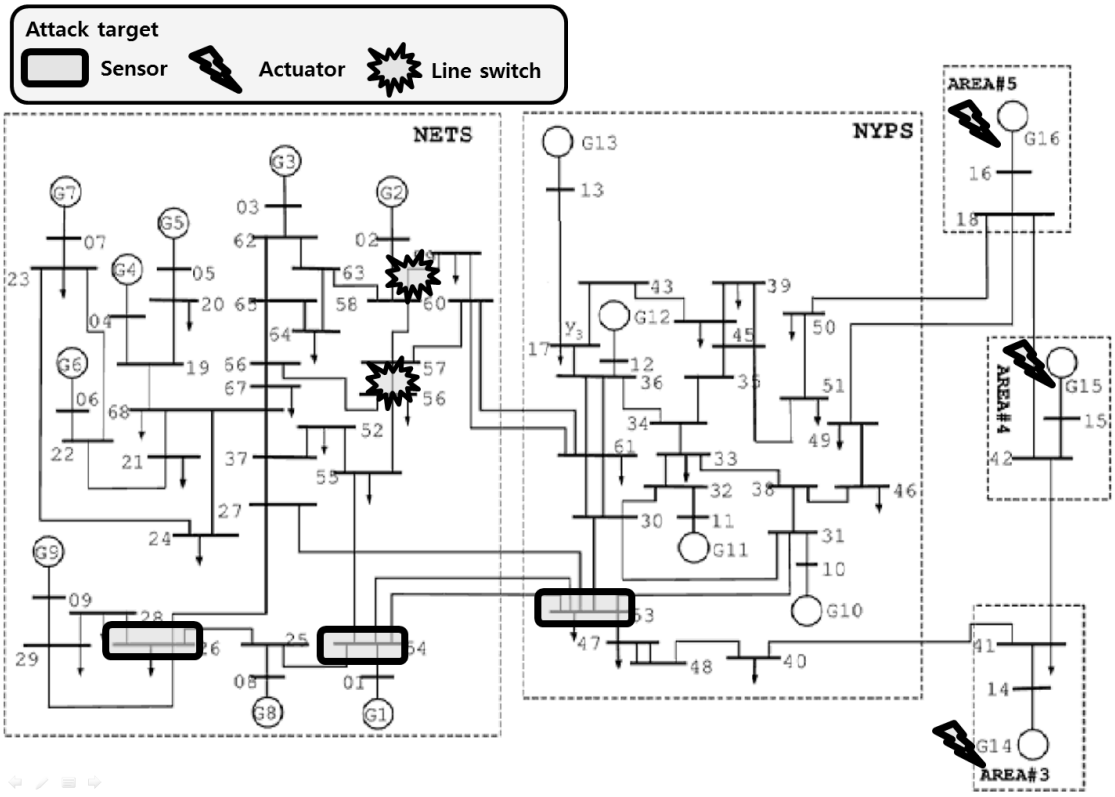

Figure 3.1: Locations of the attacks (Figure from [1]).

We consider the attack scenario where the system is under the time-varying attacks: sensor attacks $0.01\cos(0.12t)$ for $t = [0, 10)$, actuator attacks $0.1 - 0.6\sin(0.3t)$ for $t = [10, 20)$, and switching attacks for $t = [20, 30)$. For $t \geq 30$, the system would be attack-free.

The goals of case study 1 and 2 are to verify the performance of the NISME for time-varying modes with a regular mode set, and a reduced mode set, respectively.

Case study 1: Consider the following four modes.

Mode 0: Attack-free.

Mode 1: Sensors (electrical power outputs) $26, 53$ and 54 are attacked.

Mode 2: Actuators $14, 15$, and 16 are attacked.

Mode 3: Line switches $\{56, 57\}$, and $\{59, 60\}$ are attacked.

If the sizes of the estimated attack vectors are not statistically significant, mode 0 will be chosen, as described in Algorithm 2.

Remark 3.1.1 *The power system under the mode 1,2 and 3 satisfies Assumptions 2.4.1, 2.5.1, 2.5.2, 2.5.3 and uniform observability condition for all three*

modes. This is because the system is time invariant and the linearization error is $O(\|\tilde{x}_{k|k}\|^2)$ *for* $\|\tilde{x}_{k|k}\| \leq 1$. ■

Significance levels $\alpha_1 = \alpha_2 = 0.75$ are applied with corresponding chi-square values $\chi^2_3(0.75) = 4.11$. The mode probabilities and estimation results are shown in Figure 3.2 where the estimates are coincident with the true modes. Mode estimation is inaccurate near 10 sec. This is because the sizes of attack vectors are small (the second and third subfigure in Figure 3.3) during this time and thus the attack vector estimates are not considered statistically significant. Mode probabilities of mode 1 and 2 are oscillating for $t > 30$, because both modes 1 and 2 with zero attack vectors would represent attack-free mode (mode 0).

The outputs of the NISME and the real attack signals are presented in Figure 3.3. The first subfigure indicates that the state estimation errors satisfy PESp-like property. Although the frequency fluctuates due to the actuator attack $t \in [10, 20)$, its estimates are accurate. The additive sinusoidal sensor attack for $t \in [0, 10)$ is well estimated as shown in the second subfigure. The third subfigure shows the estimates and real vectors of actuator attack; the sinusoidal actuator attacks for $t \in [10, 20)$ on the control inputs. Around 10 sec, attack vector estimates are set to zero because $\hat{d}^{j=2}_{a,14}$ is not considered statistically significant.

Case study 2: There are 6 potential signal attack locations with 2 possible switching attacks, and thus we have $|\mathbb{M}^I| = 2^2$, $|\mathbb{M}^A| = 2^6$, and $|\mathbb{M}| = 2^6 \times 2^2 = 256$. We, however, could reduce the number of modes by lines 3-6 in Algorithm 5 with $\mathcal{T} = \infty$ into four; i.e., $\mathbb{M}^D_{[1,\infty]} = \{j_1, j_2, j_3, j_4\}$, where each mode associates with one $j \in \mathbb{M}^I$. All the four modes assume that sensors (electrical power outputs) $26, 53, 54$ and actuators $14, 15$, and 16 are attacked. Their assumptions on line switching attacks are described as below:
Mode j_1: There is no line switching attack.
Mode j_2: Line switches $\{56, 57\}$, and $\{59, 60\}$ are attacked.
Mode j_3: Line switch $\{56, 57\}$ is attacked.
Mode j_4: Line switch $\{59, 60\}$ is attacked.

Note that the systems under the new modes satisfy Assumptions 2.4.1, 2.5.2, and 2.5.3 and uniform observability condition as well as the rest of the assumptions in Theorem 2.5.1. True mode estimation is essential because none of the above modes is true.

We conduct the simulation for Algorithm 5, using confidence levels $\alpha_1 = \alpha_2 =$

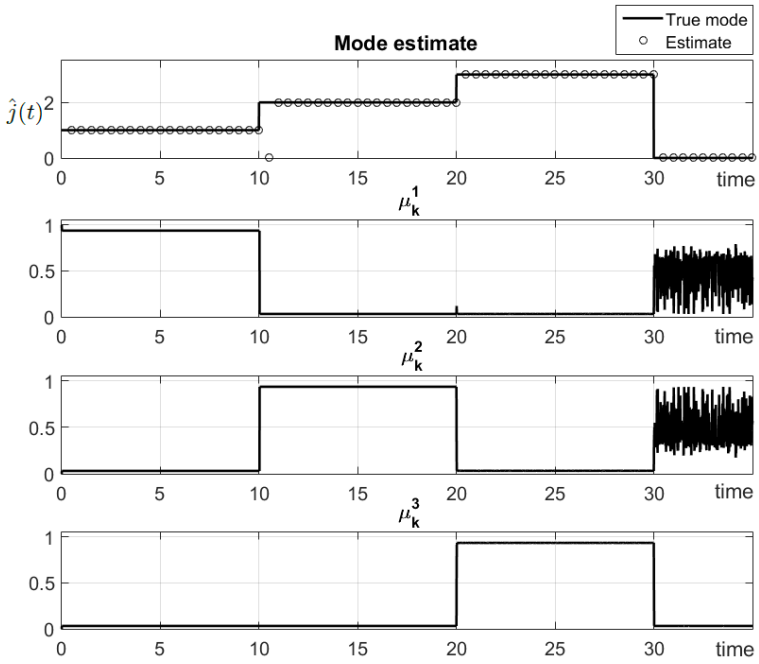

Figure 3.2: Mode estimates and probabilities of each mode.

0.8 with corresponding z-values $z(\alpha_1) = z(\alpha_2) = 1.28$. As case 1, the true modes are among modes 0 to 3, defined in case 1. This is unknown to the defender and the defender remains to consider 256 possible modes in Algorithm 4. For the presentation purpose, we project the modes other than modes 0 to 3 into mode 4; i.e., if mode 4 is chosen, mode estimation is incorrect.

The estimation results are shown in Figure 3.4, and 3.5, which are consistent with the results of the case 1 shown in Figure 3.2, and 3.3. The first subfigure in Figure 3.4 provides a true mode estimation described in Section 2.6.3. As case 1, mode estimates are erroneous near 10 sec because the sizes of attack vectors are small and thus the attack vectors are not regarded statistically significant. After 30 sec, mode probabilities oscillate between two modes in case study 1, but not in case study 2. In case study 1, two modes 1 and 2 are true with zero signal attacks, but mode 3 cannot be a true mode. In case study 2, only mode j_1 is true with zero signal attacks, but modes j_2, j_3 or j_4 cannot be a true mode.

The modes in the reduced mode set have less restrictive assumptions on attack locations than those of the original mode set, but shows similar estimation results. This simulation, thus, validates the performance of the NISME for the minimal number of modes discussed in Section 2.6.3.

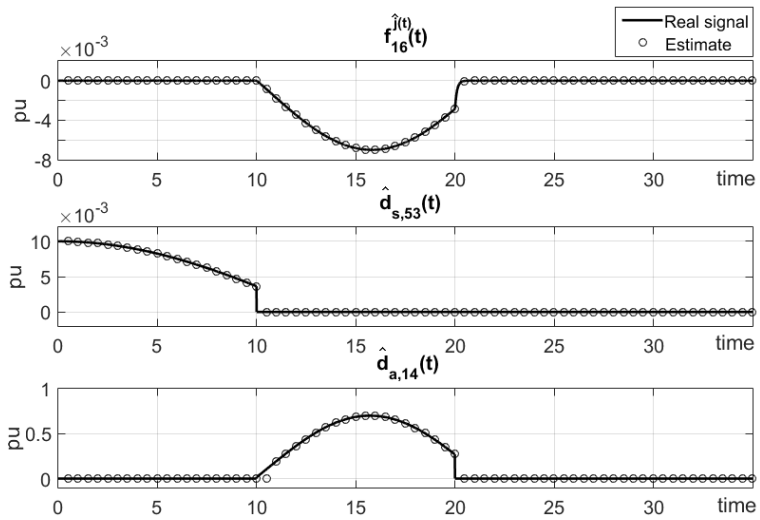

Figure 3.3: Real signals and estimated signals (top to bottom); (a) state estimation of angular frequency at bus 16; (b) sensor attack estimation of bus 53; (c) actuator attack estimation on the control input of bus 14.

3.2 Mobile robots

Figure 3.6 shows an image of the robot system. It consists of Khepera III [107] deferential drive robot mounted with KoreBot II [108] extension chip. Khepera is actuated by setting the speeds of the two wheels on its chassis. KoreBot runs OpenEmbedded Linux, which enables in-robot programming and control. The robot is equipped with three sensors: a wheel encoder, a laser range finder (LiDAR), and an indoor positioning system (IPS). The wheel encoder calculates the traveled distance of each wheel in a short period of time. Given its previous state, the traveled distance is further processed into its current position and orientation. LiDAR scans laser beams in 240 degrees of angle, and receives reflection to obtain distances from surrounding objects. IPS is powered by Vicon motion capturing system (see Figure 3.6). Multiple cameras on the roof track the positions of the reflective markers on the robot, and calculates its position.

In the real world, wheel encoder, ranger finder and positioning system (e.g. GPS, GLONASS) are common sensor settings for many ground vehicles [109, 110] in applications such as localization, obstacle avoidance, navigation, etc. Typically. positioning system serves as the primary navigation source, and wheel encoder serves as secondary source when position system is not available (e.g. in tunnels).

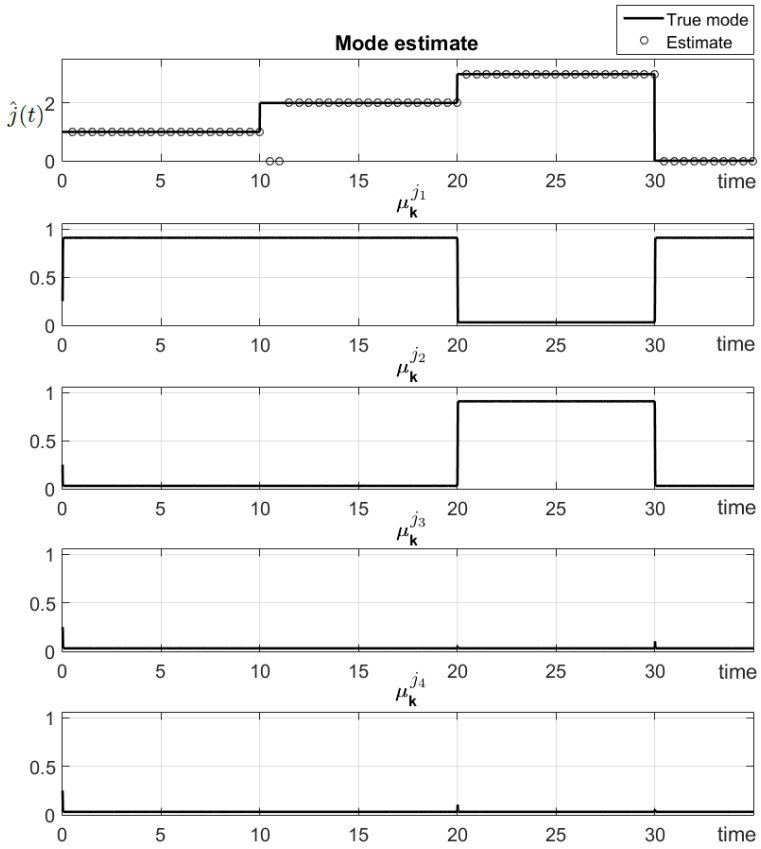

Figure 3.4: Mode estimates and probabilities of each mode with reduced mode set.

LiDAR detects nearby obstacles and redirects the robot to avoid them. We believe that our testbed reflects features of real world robots.

For comparison purpose, we use an identical path generated from RRT* for all scenarios in the experiments. In each experiment, Khepera travels from a starting point at $(0m, -1.2m)$ to a target point $(0m, 1m)$ inside a $3m \times 4m$ confined space shown in Figure 3.6, with constant 7000 speed units[1]. Three $0.8m \times 0.2m \times 0.2m$ cube-shaped obstacles reside on the ground between the starting and target location. RRT* algorithm generates a path that avoids the obstacles, and Khepera follows the path using PID ($P = 0.8, I = 0, D = 0.001$) control. We identify sensor measurement noise covariance matrix R and the process noise covariance matrix Q by referring to the data sheets of the sensors along with some empirical experiments. (refer to [111] for more systematic approaches.) NISME generates detection results under confidence level of 0.05 for actuator attacks, and 0.005 for

[1]Speed ratio 144010 units per m/s, 7000 units is approximately $0.05m/s$.

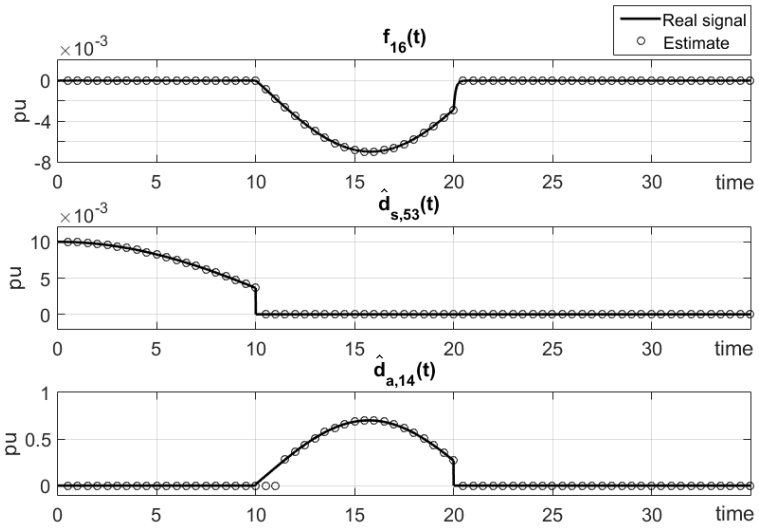

Figure 3.5: Performance of the NISE for the reduced mode set. Detailed descriptions on the other subfigures are identical to those of Figure 3.3.

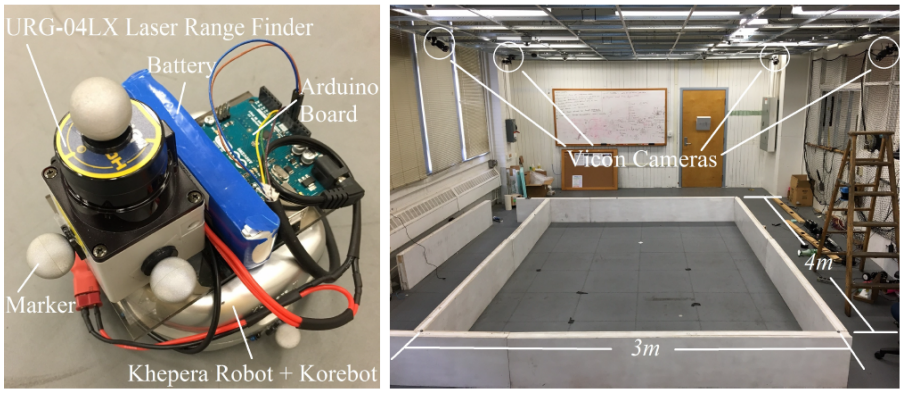

Figure 3.6: Khepera robot testbed and indoor positioning system.

sensor attacks. We choose 2 positives out of 2 windows as the decision criteria for sensor attacks, and choose 3 positives out of 6 windows as the decision criteria for actuator attacks.

Mission. Khepera has a planned path from the starting position to the target position as shown in Figure 3.7, while avoiding static obstacles. Khepera implements PID control to achieve the goal.

Attack setup. We conduct multiple attack scenarios during the mission. The complete list of the attacks are described in Table 3.1. We intend to demonstrate that NISME works well regardless of the attack channels or the target components

Table 3.1: Attack scenarios launched against Khepera mobile robot.

Attack Scenario	**Attack Scenario Description**
Wheel controller logic bomb	Logic bomb in actuator utility lib that alters the control commands to be executed
Wheel jamming	Physically jamming a particular wheel so that it will stuck
IPS logic bomb	Logic bomb in IPS data processing lib that alters the authentic positioning data
IPS spoofing	IPS signal that overpowers authentic source and sends out fake positioning data
Wheel encoder logic bomb	Logic bomb in wheel encoder data processing lib that alters its readings
LiDAR sensor blocking	Physically blocking laser ejection and reception channels in particular directions with masks
LiDAR DOS	Denial of service by cutting off LiDAR sensor wire connection with the robot
Wheel controller and IPS logic bombs	Altering both wheel control commands and IPS readings through logic bombs
LiDAR DOS and wheel encoder logic bomb	Blocking LiDAR readings and altering wheel encoder readings
IPS spoofing and LiDAR DOS	Altering IPS readings and blocking LiDAR readings
IPS and wheel encoder logic bombs	Altering both IPS and wheel encoder readings through logic bombs

in the robot platform. The attack scenarios target on different sensing or actuation workflows of the robot, and launch actuator and sensor attacks through different channels including cyber and physical channels. We inject several logic bombs into the data processing libraries of the IPS and the wheel encoder. The logic bombs can be triggered at certain time after the mission start, and continuously alter the authentic sensor readings afterwards. For instance, we can trigger the logic bomb to stealthily shift the positioning data received from IPS by a certain distance along the X axis. A logic bomb is also injected into the wheel controller library to add extra control commands to the two wheels. Wheel jamming attack is launched by physically jamming a wheel, so that the wheel stops moving. IPS spoofing attack is launched by overriding authentic IPS signals from the Vicon system and sending fake positioning data. IPS spoofing is analogous to real world GPS spoofing attacks. For LiDAR, we launch sensor attack by blocking the signal

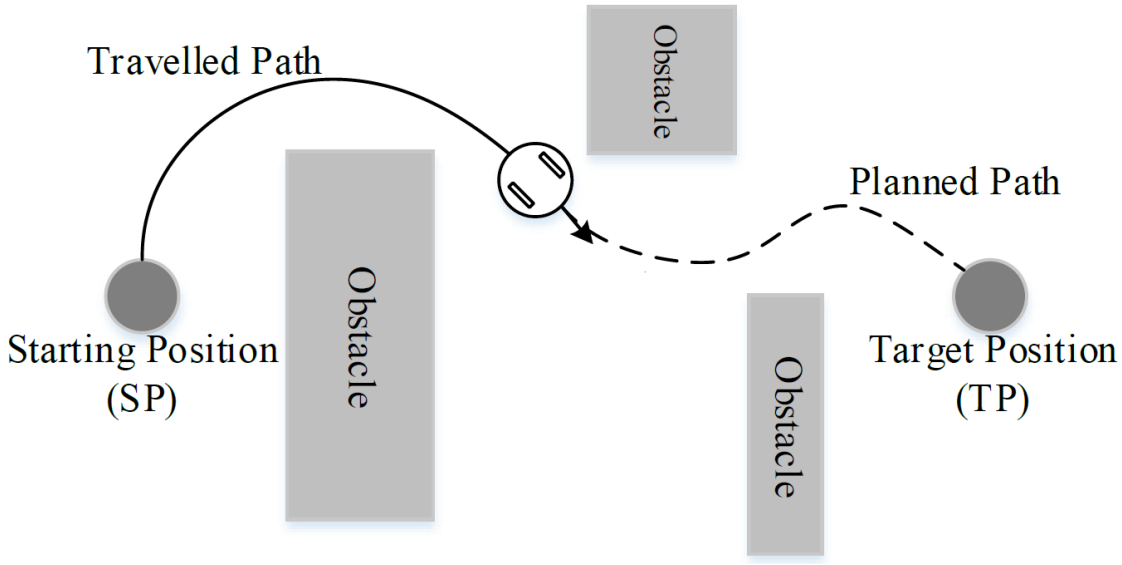

Figure 3.7: Khepera mission.

ejection and reception channel in certain directions. Besides, we launch the attack that sabotages the signal transmission by physically cutting off its wires. To evaluate the detectabiliy of NISME when multiple sensing workflows or actuation workflows are under attack, we launch several attack scenarios where several of the aforementioned attacks are combined. Table 3.3 shows quantitative information about the details of attack scenarios. In addition to attack scenarios, we also conduct 9 scenarios when the mission is finished without intrusion.

NISME aims at detecting as well as identifying attacks in robots. To evaluate the effectiveness of NISME, we define *true positive* as a time instant that 1) raise an alarm if the robot is under attack, and 2) identify the correct attack target case; i.e., which sensor or actuation workflow(s) are attacked. Otherwise, positive detection result is considered as *false positive*. *False negative* is defined as a time instant when NISME does not raise alarm when any workflow is under attack. If all workflows are free of attacks and NISME does not raise any alarm, the time instant is referred to as *true negative*. The detection result column in Table 3.3 shows identification of attack types and attack target case for different scenarios. Note that some scenarios (e.g. #7) are not provided with quantitative description on the attack vector because some experiments are difficult to describe. From the 11 attack scenarios, we observe that both types of attacks launched from different channels can be successfully detected and identified. Scenario #1, #2 and #8

involves actuator attacks launched from different channels, and #8 is also mixed with sensor attacks. Scenario #3 and #4, #6 and #7 are two pairs of attacks launched to the same sensors. Scenario #8, #9 and #10 are mixed sensor attacks where two sensors are compromised.

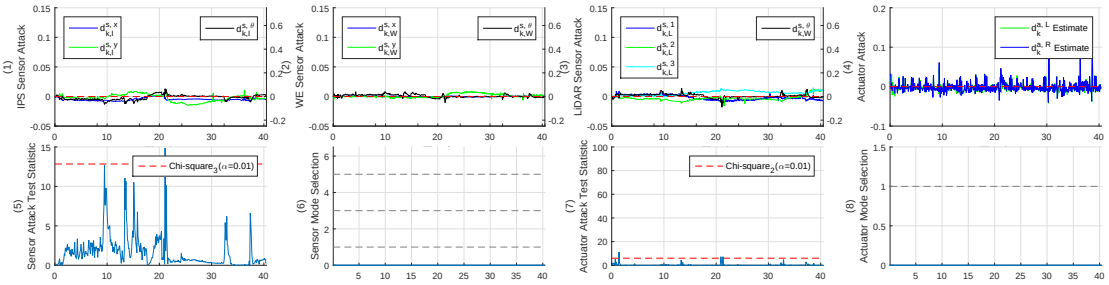

Figure 3.8: No attack scenario. The eight plots in each subfigure are: (1) estimated sensor attack vector on IPS; (2) estimated sensor attack vector on wheel encoder; (3) estimated sensor attack vector on LiDAR; (4) estimated actuator attack vector for the wheels; (5) sensor attack Chi-square hypothesis test statistic and threshold under confidence level $\alpha = 0.005$; (6) sensor attack target case selection; (7) actuator attack Chi-square hypothesis test statistic and threshold under confidence level $\alpha = 0.05$; (8) actuator attack target case selection.

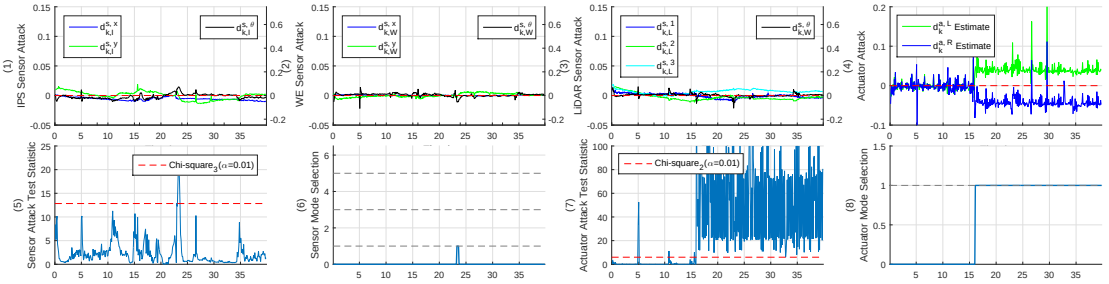

Figure 3.9: Attack scenario #1: wheel controller logic bomb.

Detection Results. For the ease of presenting classification results, Table 3.2 defines the attack target cases for actuator and sensor attacks. Note that sensor attack target case S_4, S_5 and S_6 represent cases when multiple sensor readings are corrupted and only one sensor returns uncorrupted value. In each iteration we also calculate the deviation between reference sensor readings and estimated sensor readings using state estimates so that we can see the estimation for all sensors.

Figures 3.8-3.12 present graphical details of the detection results for several attack scenarios. Each figure includes eight plots that elaborate the outputs of

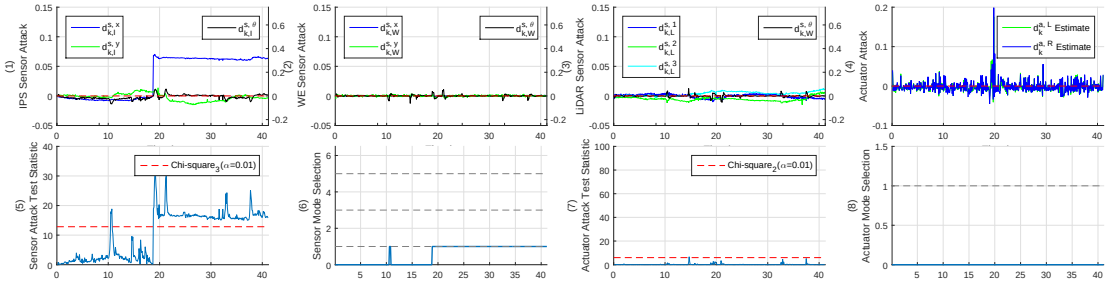

Figure 3.10: Attack scenario #3: IPS logic bomb.

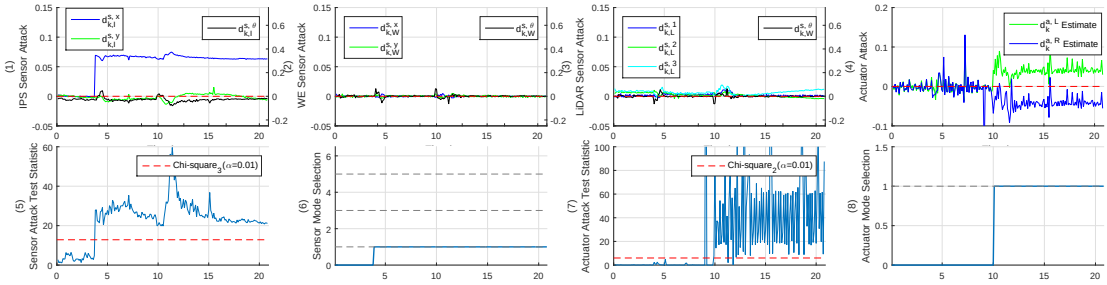

Figure 3.11: Attack scenario #8: Wheel controller and IPS logic bomb.

NISME in each experiment: 1) estimated sensor attack vector on IPS; 2) estimated sensor attack vector on wheel encoder; 3) estimated sensor attack vector on LiDAR; 4) estimated actuator attack vector for the wheels; 5) sensor attack Chi-square hypothesis test statistic and threshold under confidence level $\alpha = 0.005$; 6) sensor attack target case selection; 7) actuator attack Chi-square hypothesis test statistic and threshold under confidence level $\alpha = 0.05$; 8) actuator attack target case selection. Figure 3.8 shows the detection output when there is neither actuator attack nor sensor attack. Estimation results in plot 1-4 show nearly zero estimated attack vectors with noises. The Chi-square test statistics shown in plot 5 and 7 indicate both actuator and sensor attack remain under predetermined threshold, except some occasional spikes caused by noises. After the sliding window filtering, plot 6 and 8 indicates an attack silence. Figure 3.11 shows a scenarios when wheel controller control commands and IPS sensor readings are tampered by logic bombs at different time instants. Around $4s$, sensor attack vector estimates on the X axis of IPS readings surges (plot 1). Accordingly, sensor attack test statistic surges above the threshold (plot 5), and sensor attack target case selection (plot 6) indicates that the robot is under IPS sensor attack. Around $10s$, actuator attack vector estimates on left and right wheel significantly deviate from 0. Accordingly,

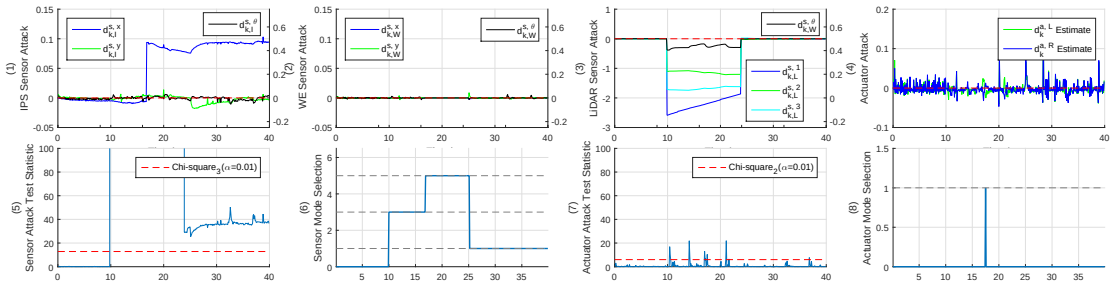

Figure 3.12: Attack scenario #10: IPS spoofing and LiDAR DOS.

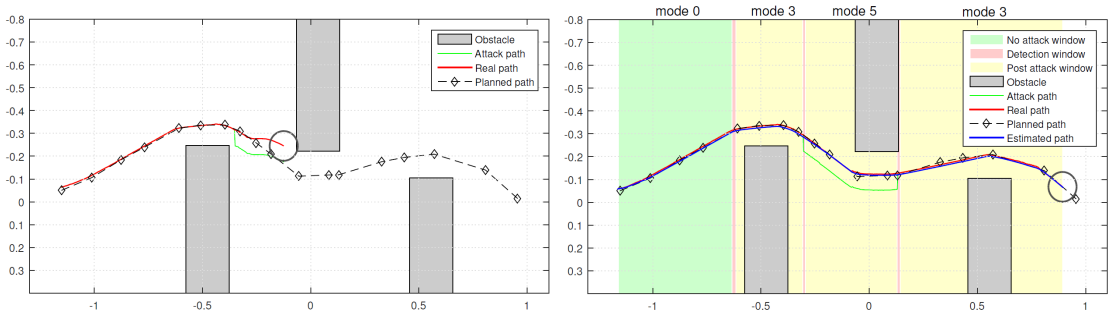

Figure 3.13: Mission execution under attack scenario #11. Khepera travels from the start location (left side) to a target location (right side). The green background indicates the time window when there is no attack. The red background indicates the time window after an attack is launched/revoked and before RIDS correctly identifies the change. The yellow background indicates the time window under attack and RIDS correctly identifies the attack.

we notice an oscillating surge over the threshold for actuator attack (plot 7), and actuator attack target case selection (plot 8) indicates that the robot is under actuator attack. During the experiment, both sensor attack estimates for wheel encoder and LiDAR remain silent. Figure 3.12 shows a scenario where attacks against multiple sensors are launched/revoked at four different time instants. We observe that the detection results are highly consistent with the scenario design.

We examine the false positive and false negative time instants occurred in the experiments. Majority of false classifications are introduced by the sliding window for the purpose of denoising. False positives and false negatives are inevitable at the edge when attacks become active or revoked, and the choice of window size and decision criteria determine the number of false classifications. For sensor attack false positives, we observe only a small portion is caused by attack target case

[1]False positive rate and false negative rate (%).

Table 3.2: Sensor and actuator attack target case definition.

Sensor Attack Target Case #	**Robot Attack Status**
S_0	under no sensor attack
S_1	under IPS sensor attack
S_2	under wheel encoder sensor attack
S_3	under LiDAR sensor attack
S_4	under wheel encoder and LiDAR sensor attack
S_5	under IPS and LiDAR sensor attack
S_6	under IPS and wheel encoder sensor attack
ActuatorAttack Target Case #	**Robot Attack Status**
A_0	under no actuator attack
A_1	under actuator attack

selection errors, while majority is caused by bogus test statistics increases. The average false positive rate and false negative rate are 0.86% and 0.97%, respectively. Therefore, the NISME is considered effective in detecting and identifying both actuator attacks and sensor attacks targeted in our testbed.

Detection Delay. Detection delay indicates the time between when specific attack is launched/revoked, and when NISME captures the change. Theoretically, in each control iteration, attack vectors can be revealed in the very next iteration after launch time from NISE. However, we add a sliding window in the decision maker to eliminate noise impact. Hence, detection delays will depend on the decision making algorithm and parameter choice. In our experiment, we choose sliding window as the detection making algorithm. Specifically, we choose 2/2 and 3/6 as the decision criteria and sliding window size. The detection delay for each attack scenario is shown in Table 3.3. We observe that the detection delays are quite small. Specifically, average detection delay for sensor attacks is $0.35s$, and the counterpart for actuator attacks is $0.61s$. The average delays are consistent with our parameter selection for actuator and sensor attacks. Through our analysis on the detection statistics, we notice that NISME raises alarm mostly in the next iteration after attack occurs. Most delays are incurred by the sliding window decision making.

Once the magnitude of an attack exceed predetermined threshold, the maximal detection delay is a constant multiple of control iterations. The frequency of control iteration is determined by hardware configurations (e.g. CPU frequency) and

Table 3.3: Attack scenarios and detection results from NISME.

#	Attack Scenario	Launch Time	Attack Type (Channel)	Attack Description	Detection Delay	FPR /FNR[2]
1	Wheel controller logic bomb	16.0	Actuator (cyber)	-6000 units on v_L +6000 units on v_R	0.49	A: 0/0.83 S: 1/-
2	Wheel jamming	5.3	Actuator (physical)	0 unit on v_L	0.76	A: 0/3.1 S: 0/-
3	IPS logic bomb	19.0	Sensor (cyber)	shift $+0.07m$ on X	0.30	A: 0/- S: 1.6/0.24
4	IPS spoofing	26.0	Sensor (physical)	shift $-0.1m$ on X	0.24	A: 2.24/- S: 1.55/1.39
5	Wheel encoder logic bomb	16.0	Sensor (cyber)	increment 100 steps on left wheel encoder	0.43	A: 1.4/- S: 0/0.45
6	LiDAR DOS	0.0	Sensor (physical)	received distance reading is $0m$ in each direction	0.23	A: 0/- S: 0/0
7	LiDAR sensor blocking	7.0	Sensor (physical)	received distance reading to the left wall is erroneous	0.55	A: 0.22/- S: 0/0.80
8	Wheel controller & IPS logic bomb	W: 10.0 I: 3.8	Sensor &Actuator (cyber)	$\mp$6000 units on v_L, v_R shift $+0.07m$ on X	W: 0.59 I: 0.50	A: 0/1.8 S: 0/0.24
9	LiDAR DOS & wheel encoder logic bomb	W: 16.0 L: 25.0	Sensor (cyber &physical)	increment 100 steps on left wheel $0m$ in each direction from LiDAR	W: 0.43 L: 0.29	A: 0/- S: 0.48/0.72
10	IPS spoofing & LiDAR DOS	L: 10.0 I: 17.0 L: 25.0	Sensor (physical)	$0m$ in each direction from LiDAR shift $+0.07m$ on X LiDAR readings are restored to normal	L: 0.36 I: 0.29 L: 0.30	A: 0.25/- S: 0.25/0.58
11	IPS & wheel encoder logic bomb	W: 10.0 I: 28.0	Sensor (cyber)	increment 100 steps on left wheel shift $+0.1m$ on X	W: 0.33 I: 0.31	A: 0/- S: 0.25/0.33

control algorithm design, which is chosen to meet the specifications of robots and operational needs. Fast moving robots should have higher frequency of control cycles to ensure timely sensor data processing and actuation. For instance, to operate in a harsh field environment under a relatively high speed, Boston Dynamics Big-

Dog [112] is designed with a frequency of 200HZ to facilitate robot balancing and steering. In our testbed, the frequency of the control iteration is 100HZ. We observe that the detection is much earlier than the collision. Security administrative has plenty of time for attack response. Moreover, since certain safety policies are usually enforced into a mission (e.g. safety minimum distance from obstacles for ground vehicles, safety minimum altitude for UAVs), we believe that our NISME can quickly detect attacks before they cause significant damage to the robot or the environment.

Attack Vector Quantification. Actuator attack and sensor attack vector estimation provides quantitative information of the attacks, which can assist security administrative for further diagnosis and attack response. For instance, after sensor attack detection in scenario #9, IPS sensor attack estimate on X axis is $+0.069m$ with standard deviation of $\pm 0.002m$. Average error between estimated vector and the ground truth ($+0.07m$) is 1.91%. After actuator attack detection, average actuator attack estimates on the left wheel and right wheel are -5975.4 ± 1188 units and $+5892.4 \pm 1091$ units, respectively. Average error between estimated vector and the ground truth (∓ 6000 units) are 0.41% and 1.79%, respectively. We observe that the estimation results are fairly accurate for both actuator and sensor attack vector estimation.

3.3 Connected vehicles

We consider a multi vehicle system, where each vehicle equips GPS, wheel encoders, and IMU (Inertial Measurement Unit) for navigation, and LiDAR (Light Detection And Ranging sensor) to observe the states of neighboring vehicles (see Figure 3.14). A vehicular ad hoc network in Figure 3.15 is established to connect nearby vehicles in a decentralized and self-organizing manner.

We consider adversaries that can launch active attacks on the vehicles in a connected vehicle network in order to deviate the vehicle from its normal operation. The adversaries can observe real-time vehicle states and has knowledge about vehicle sensing, actuation, and computing systems. They are capable of launching sensor attacks or actuator attacks through different channels, including physical damage (e.g. jamming wheels), signal interference (e.g. GPS spoofing), or cyber breach (e.g. root-kit) on one or multiple vehicles.

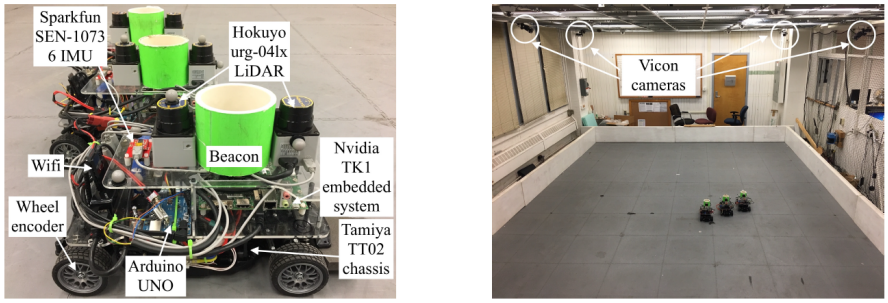

Figure 3.14: Scaled autonomous vehicle testbed and indoor positioning system.

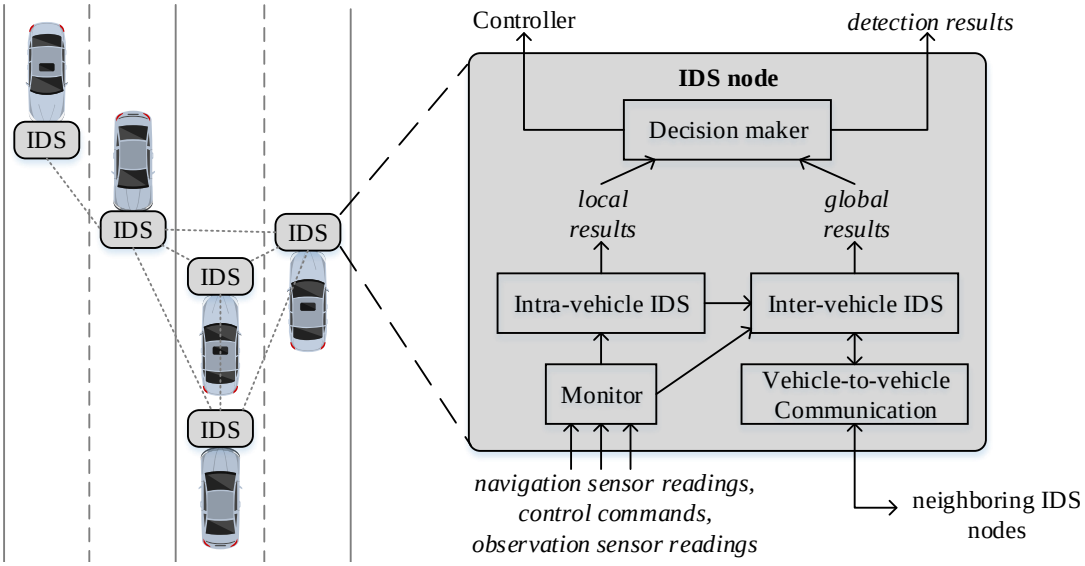

Figure 3.15: Vehicle collaborative intrusion detection system overview.

The VCIDS (Vehicle Collaborative Intrusion Detection System) consists of the intra-vehicle IDS (Intrusion Detection System) and inter-vehicle IDS as shown in Figure 3.15. In particular, the intra-vehicle IDS module is dedicated for the purpose of detecting local sensor and actuator attacks, as well as generate state estimates using local data. The intra-vehicle IDS applies the NISME on the local data collected from the monitor. The inter-vehicle IDS is dedicated to confirm the attacks detected from the intra-vehicle IDS, and identify a boarder range of attacks by the NISME. The key audit data source is the observation sensor readings from nearby vehicles. Once the new state estimates is generated, a vehicle can estimate the state of nearby vehicles within the range of its observation sensors. After that, each vehicle receives the state estimates of itself from nearby vehicles. Note that the number of observed vehicles can be different from the number of received state

estimates. Then, the received state estimates are treated as sensor readings from external sources and fed into the multi-mode estimation algorithm along with the clean sensor readings identified from the intra-vehicle IDS. Finally, the detection results for the received observations are broadcasted, and receive the corresponding results for decision making.

We evaluate the VCIDS on the scaled autonomous vehicle testbed against various attacks and demonstrate its security capabilities. We intend to answer two research questions for the detection system: 1) what benefits does the VCIDS offer in terms of security capabilities? 2) To what extent does the VCIDS influence the detection performance, i.e., effectiveness and efficiency? We present the detection results generated from the VCIDS nodes and compare the detection results between deploying intra-vehicle IDS only and deploying both IDSs.

The three vehicles in the testbed travel in the indoor environment. For the ease of presentation, we label the three vehicles with fixed numbering. In each experiment, vehicle 1 and vehicle 2 circle around the environment in a predefined two-lane road with an identical constant speed of $6cm/s$ as shown in Figure 3.16. Vehicle 3 stays on the roadside without moving, but all onboard sensors are working. During the mission, the three vehicles communicate with each other through V2V communication and collaboratively detect attacks.

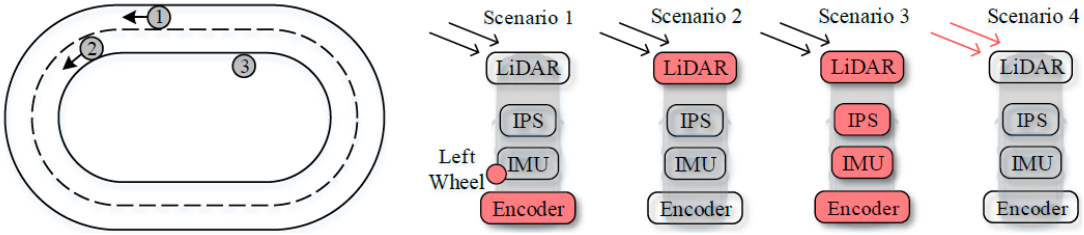

Figure 3.16: Scaled autonomous vehicle execution in the indoor environment. Attack scenarios. Scenario 1: Encoder logic bomb and left wheel jamming. Scenario 2: LiDAR driver logic bomb. Scenario 3: System hijacking. Scenario 4: Rogue nodes.

Attack Setup. To demonstrate the effectiveness of our detection system, we consider the following four attack scenarios where attacks are launched on different targets. The attack scenarios are conducted independently with each other.

Wheel encoder logic bomb & wheel jamming. The attack is launched by replacing the wheel encoder sensor data processing library with a customized library

in vehicle 1. Instead of returning states obtained from motion of the wheel shafts, the customized library returns the sensor readings with a constant sensor attack vector that shifts the vehicle by $-10cm$ on the X axis. A plastic stick is placed in the left rear wheel of vehicle 1. The stick adds friction in the wheel and slows down the movement of the wheel. The actuator attack adds a inconstant attack vector on the wheel.

LiDAR driver logic bomb. Analogous to the wheel encoder sensor logic bomb attack, we replace the driver program of the LiDAR observation sensor driver with a customized sensor driver of vehicle 1. The customized driver alters the relative distances and angle measurements of nearby vehicles.

System hijacking. For advanced attackers, it has been demonstrated to be possible that attackers can hijack into the vehicle system. Hence, it is possible that attackers could corrupt all sensor readings simultaneously. In order to avoid detection, an advanced attacker would try to achieve their attack goal while avoiding the detection. To do this, the attacker could modify all sensor readings in a consistent manner. For instance, an attacker could shift all sensor readings on Y axis by $+10cm$. During the intra-vehicle detection phase, the multi-mode estimation algorithm does not have a clean sensor as the reference sensor. Moreover, since the sensor readings are corrupted consistently, the hypothesis tests would not generate positive results due to the lack of significant deviation. Here we consider an advanced attacker who carefully crafts all sensor data in vehicle 1 and makes them consistent with each other.

Rogue nodes. Attackers can setup rogue nodes that broadcast fake messages to nearby vehicles that are intended to cause wrong decision making for vehicles. In this scenario, we assume that a rogue node is set up by the roadside which intend to broadcast fake observations. The rogue node broadcasts large amount of fake observations of vehicle 1 that contain shifted observations.

Detection Results. In order to demonstrate the security capability of the collaborative detection system over a standalone intra-vehicle IDS, We compare the detection results generated from the intra-vehicle IDS and the complete VCIDS. Table 3.4 shows the detection results generated from the four attack scenarios we launched in the testbed. We observe that the intra-vehicle IDS can only detect the first attack scenario when a subset of navigation sensors are under attack. On the contrary, the VCIDS detects all attack scenarios. When the navigation sensor is

under attack (Scenario 2), state estimation for nearby vehicles are corrupted. When vehicle 2 and vehicle 3 receives the corrupted observation from vehicle 1, their inter-vehicle IDSs raise sensor attack alarm and send the results back to vehicle 1. When all sensor readings in vehicle 1 are corrupted consistently (Scenario 3), the intra-vehicle IDS of vehicle 1 does not raise alarm. However, the observations from vehicle 2 and vehicle 3 used in the inter-vehicle IDS of vehicle 1 can discover such attacks. Under rogue nodes attack (scenario 4) when fake nodes are broadcasting erroneous observations, the inter-vehicle IDS can detect the attack.

Table 3.4: Attack scenarios and corresponding detection results from intra-vehicle IDS and final results of the VCIDS.

Attack Scenario	**Attack Type (Channel)**	**Detected by Intra-vehicle IDS**	**Detected by VCIDS**
Wheel encoder logic bomb+wheel jamming	sensor+actuator (cyber+physical)	Yes	Yes
LiDAR driver logic bomb	sensor (cyber)	No	Yes
System hijacking	sensor (cyber)	No	Yes
Rogue nodes	sensor (cyber)	No	Yes

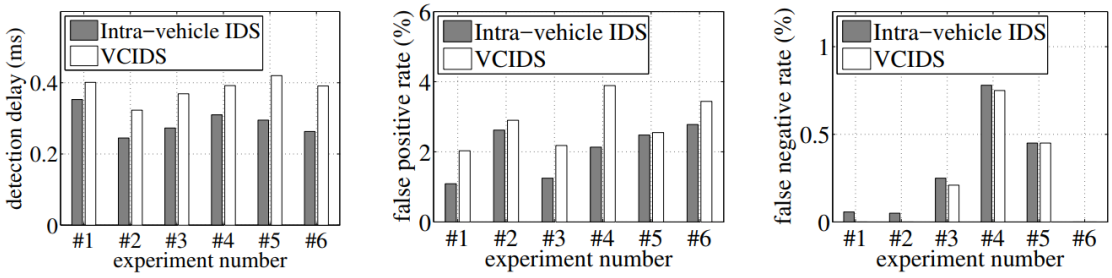

Figure 3.17: Detection performance comparison between results from intra-vehicle IDS and VCIDS.

To investigate the detection performance in terms of the detection delay and detection accuracy, we conduct some experiments launched with attack cases that can be detected by the standalone IDS. In the detection results, a false positive is defined as a time instant that raises alarm for an uncorrupted sensor, and a false negative is defined as a time instant that alarms is not raised when a sensor is corrupted. Figure 3.17 shows the comparison for detection delay, false positive

rate and false negative rate. We notice that detection delays increase since the VCIDS requires more steps after the intra-vehicle detection. We also notice a slight increase on the false positive rate and a decrease on the false negative rate. Both rates remain under 4%.

Chapter 4
Cyber-physical security: Attack-resilient machine learning

4.1 Introduction

Machine learning is increasingly used in cyber-physical systems (CPS) for a broad area of applications such as image recognition in self-driving vehicles [113], control of energy systems [114], and healthcare systems [115]. These data-driven techniques are well suited for complex systems whose models are challenging to obtain using first-principles. However, machine learning systems are threatened by cyber-attacks. It has been shown in [116, 117] that structural properties of machine learning can be used to generate adversarial samples in image recognition. Paper [118] shows that, regardless of their structures, machine learning systems including traffic sign learning systems fail under black-box attacks, which inject small perturbation to legitimate samples. Protecting machine learning systems from cyber-attacks is imperative.

There is a wealth of literature on cyber-attack detection of CPS. The fundamental limitations of attack detection in linear systems are identified in [31]. Paper [119] derives reachable regions of internal states driven by sensor attacks where the attacks can bypass a Chi-square test based on Kalman filter. A number of attack detectors against signal attacks have been proposed. In particular, attack detection problems for deterministic linear systems are formulated as ℓ_0/ℓ_∞ optimization problems [31, 32], which are NP-hard in general. To address the computational challenges, paper [32] proposes convex relaxations of the optimiza-

tion problems. Paper [33] studies attack-resilient state estimation in the presence of modeling errors, and identifies a bound of state estimation errors induced by modeling errors. Paper [91] designs an attack-resilient estimator for stochastic linear systems in the presence of sensor attacks, actuator attacks, and switching attacks. The papers mentioned above focus on linear systems, and Chapter 2 extends the results in paper [91] to accommodate a class of nonlinear systems. All the aforementioned papers presume knowledge on dynamic system models. This key assumption is relaxed in the current chapter using data-driven techniques.

Chapter organization. Section 4.2 formulates the attack-resilient machine learning problem for a class of nonlinear systems. Notations, notions, and Gaussian process are introduced in Section 4.3 as preliminaries. In Section 4.4, we design the attack-resilient Gaussian process regression to address the problem. Section 4.5 analyzes average case learning errors of the proposed regression algorithm. In Section 4.6, numerical simulations on the IEEE 68 bus test system show the performance of the proposed algorithm.

4.2 Problem formulation

Consider the nonlinear stochastic system

$$\begin{aligned} x_k &= f(x_{k-1}, u_{k-1} + d_{a,k-1}) + w_{k-1} \\ y_k &= C_k x_k + d_{s,k} + v_k \end{aligned} \tag{4.1}$$

where $x_k \in \mathbb{R}^n$, $y_k \in \mathbb{R}^m$, $u_k \in \mathbb{R}^a$, $d_{a,k} \in \mathbb{R}^a$ and $d_{s,k} \in \mathbb{R}^m$ are state, output, input, actuator attack vector, and sensor attack vector, respectively. We assume that noise vectors $w_k \in \mathbb{R}^n$ and $v_k \in \mathbb{R}^m$ are independent and identically distributed zero-mean Gaussian, with their covariance matrices $Q = \mathbb{E}[w_k w_k^T]$ and $R = \mathbb{E}[v_k v_k^T]$. We assume that R is a diagonal matrix.

Attack model. Signal injection attacks are comprised of signal magnitude attacks; i.e., the attacker injects attack signals, and signal location attacks; i.e., the attacker chooses targeted sensors and actuators. Signal injection attacks are modeled by actuator attack $d_{a,k}$ and sensor attack $d_{s,k}$ where zero value of either attack vector indicates that the corresponding actuator or sensor is free of attack, and a non-zero value indicates the magnitude of the attack.

Knowledge of the defender. System function f in (4.1) is unknown to the defender while output matrix C_k is known. The defender is accessible to input u_k and output y_k but is unaware of the attack vectors $d_{a,k}$ and $d_{s,k}$, as well as which actuators/sensors are under attacks. Noise vectors w_k, v_k and autocovariance Q are unknown but R is known.

Objective. The defender aims to recursively estimate internal state x_k, attack vectors $d_{a,k}$, $d_{s,k}$ and system function f in the presence of sensor attacks and actuator attacks.

4.3 Preliminaries

This section summarizes the notations and notions used throughout the chapter. It also discusses classic GPR following the presentation in [2].

4.3.1 Notations and notions

Hat notation over a variable denotes an estimate of the variable. In particular, $\hat{x}_{k|k-1}$ is a predicted state (an estimate without the current output); $\hat{x}_k$ is a state estimate (an estimate with the current output); $\hat{d}_k$ is an estimate of attack vector of d_k; and $\hat{f}$ is an approximation of function f. Also, $\tilde{a}_k \triangleq a_k - \hat{a}_k$ denotes the estimation error and $P_k^a \triangleq \mathbb{E}[\tilde{a}_k \tilde{a}_k^T]$ denotes the error covariance of a_k. Let $dim(v)$ denote the dimension of vector v. Gaussian distribution is denoted by $\mathcal{N}(\mu, \Sigma)$, where μ is mean and Σ is variance.

Definition 4.3.1 *(Definition 6.1 in [2]) Let $\mathcal{H}$ be a Hilbert space of real functions f defined on $\mathcal{X}$. Then, $\mathcal{H}$ is called a reproducing kernel Hilbert space (RKHS) endowed with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ if there exists a unique function g such that for every $x \in \mathcal{X}$, $g(x, x')$ as a function of x' belongs to $\mathcal{H}$, and g has the reproducing property $\langle f(\cdot), g(\cdot, x) \rangle_{\mathcal{H}} = f(x)$.* ■

In the above definition, function g is called kernel, and $\|f\|_{\mathcal{H}} \triangleq \sqrt{\langle f, f \rangle_{\mathcal{H}}}$ is a norm induced by the inner product. An example of RKHS is the space of bandlimited continuous functions $\mathcal{H} = \{f \in C(\mathbb{R}) | supp(F) \subset [-a, a]\}$ where $C(\mathbb{R})$ is the set of continuous functions, F is the Fourier transform of f, and a is the band limit. Corresponding kernel is $g(x, x') = \frac{a}{\pi} \text{sinc}(a(x - x'))$ and inner product is defined by $\langle f, g \rangle \triangleq \int f(x) \bar{g}(x) dx$.

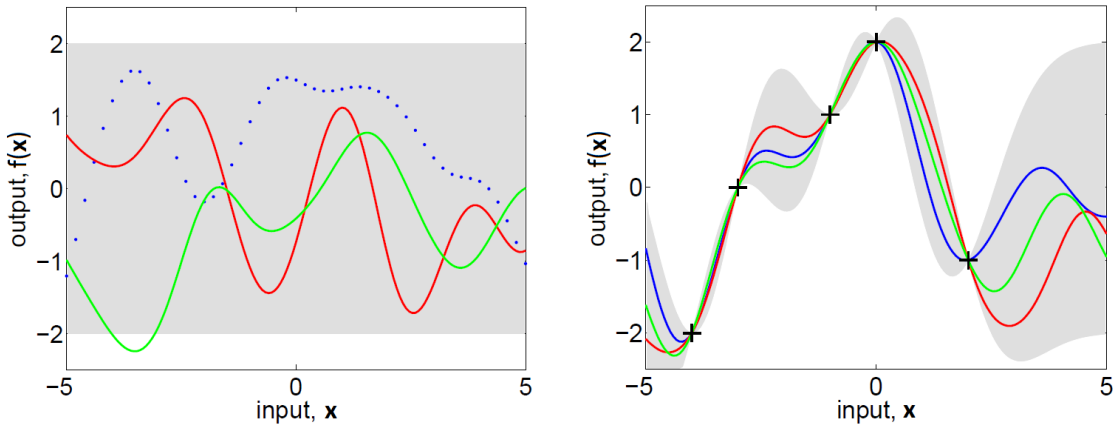

Figure 4.1: Illustrative example of GPR borrowed from [2]. Left: Random functions drawn by GP prior and actual outputs $f(\mathbf{x})$ (dots). Right: Random functions drawn by GP posterior from 5 noise-free observations indicated by +. Shaded area is point-wise 95% confidence region.

4.3.2 Gaussian process regression

GPR is a regression algorithm by GP implementation and has several advantages over parametric regression techniques such as linear and nonlinear regressions. First of all, it is non-parametric regression, and does not require prior knowledge of target functions such as structural properties. Second, GP prediction interpolates missing observations, providing empirical confidence intervals. Moreover, it is easy for users to provide interpretations of data correlations by choosing a kernel function.

Illustrative example. Consider simple noise-free equation $\mathbf{z} = f(\mathbf{x})$, where function $f : \mathbb{R} \to \mathbb{R}$ is a continuous regression function. We would like to approximate f and obtain test output $f(\mathbf{x}_*)$, given test input $\mathbf{x}_*$. GPR approximation consists of two steps; prior prediction and posterior prediction. In the prior prediction step, no information is available and the regression function is assumed to be GP. In other words, any finite collection of outputs of f are multi-variate Gaussian. In the left subfigure in Figure 4.1, the shaded area is point-wise 95% confidence region. Two solid lines are random functions drawn by GP prior, and we could draw infinitely many functions. Once we observe several pairs of inputs and outputs of f, GP prior is updated to GP posterior by incorporating the observations. In particular, the regression function must pass by the observed pairs because the

observations are noise-free. Moreover, if a test input is close enough to one of the observed inputs, the output of the test input is expected to be close enough to the corresponding observed output, having small output uncertainty. Posterior uncertainties are smaller than prior uncertainties. Under this information update, we could redraw observed points +, and point-wise 95% confidence region in the right subfigure in Figure 4.1. The solid lines in the subfigure are random functions drawn by GP posterior. Information update can be seen as a procedure to reject functions (drawn from prior) which do not agree with the observations.

Gaussian process regression. Consider the regression model

$$\mathbf{z} = f(\mathbf{x}) + w \tag{4.2}$$

with input $\mathbf{x} \in \mathbb{R}^n$ and scalar output $\mathbf{z} \in \mathbb{R}$ where $w \in \mathbb{R}$ is zero-mean Gaussian noise with variance σ^2. We will extend the result to the case with vector output $\mathbf{z} \in \mathbb{R}^m$ later.

We are going to approximate function f in (4.2), given a set of input-output observations. A pair $\mathbf{x}_i, \mathbf{z}_i$ of input-output observation is called training data. A set $D = \langle \mathbf{X}, \mathbf{Z} \rangle$ of training data is given where

$$\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_i, \cdots, \mathbf{x}_N],\ \mathbf{Z} = [\mathbf{z}_1, \cdots, \mathbf{z}_i, \cdots, \mathbf{z}_N]$$

and N is the number of the training data pairs, and index i represents i^{th} training data. GPR aims to approximate function f in (4.2) by utilizing the training data set D under the assumption that f is a zero-mean GP. Furthermore, given test input $\mathbf{x}_*$, we desire to estimate test output $\mathbf{z}_* = f(\mathbf{x}_*)$ using the function approximation.

Definition 4.3.2 *Stochastic process f is Gaussian if $[f(\boldsymbol{x}_1), \cdots, f(\boldsymbol{x}_N)]^T$ is a multivariate Gaussian random variable, for any finite set of points $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]$ (p.540 in [120]).*

Remark 4.3.1 *Multivariate Gaussian assumption on function f is used to interpolate missing observations, providing empirical confidence intervals. Although function f is assumed to be zero-mean in the prior GP prediction, a posterior distribution will not be zero-mean and the zero-mean prior assumption will be overwhelmed by a set of large training data.* ■

Under the GP assumption, $[f(\mathbf{x}_1), \cdots, f(\mathbf{x}_N)]^T$ is multivariate Gaussian and we denote its covariance matrix by (kernel matrix) $G(\mathbf{X}, \mathbf{X})$, where (i, i') element of G is denoted by kernel $g(x_i, x_{i'})$. Kernel represents a similarity between the outputs. Please refer to Table 4.1 in [2] for commonly used kernel functions, including linear, Gaussian, exponential, etc.

Since $[f(\mathbf{x}_1), \cdots, f(\mathbf{x}_N)]^T$ and $[w_1, \cdots, w_N]$ are both zero-mean Gaussian, a collection of outputs $\mathbf{Z}$ follows zero-mean Gaussian distribution

$$p(\mathbf{Z}) = \mathcal{N}(0, G(\mathbf{X}, \mathbf{X}) + \sigma^2 I).$$

Given test input $\mathbf{x}_*$, the training outputs $\mathbf{Z}$ and the test output $\mathbf{z}_*$ are jointly Gaussian:

$$\begin{bmatrix} \mathbf{Z} \\ \mathbf{z}_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} G(\mathbf{X}, \mathbf{X}) + \sigma^2 I & G(\mathbf{X}, \mathbf{x}_*) \\ G(\mathbf{x}_*, \mathbf{X}) & g(\mathbf{x}_*, \mathbf{x}_*) + \sigma^2 \end{bmatrix}\right).$$

According to p.200 in [2], the Gaussian predictive distribution over test output $\mathbf{z}_*$ has mean

$$\mu(\mathbf{x}_*, D) = g_*^T (G(\mathbf{X}, \mathbf{X}) + \sigma^2 I)^{-1} \mathbf{Z} \tag{4.3}$$

and variance

$$\Sigma(\mathbf{x}_*, D) = g(\mathbf{x}_*, \mathbf{x}_*) - g_*^T (G(\mathbf{X}, \mathbf{X}) + \sigma^2 I)^{-1} g_* + \sigma^2 \tag{4.4}$$

where $g_* = G(\mathbf{X}, \mathbf{x}_*)$. We call them as GPR mean and GPR variance, respectively. GPR mean (4.3) can be seen as a weight average of training outputs $\mathbf{z}_i$ with corresponding weight vector $g_*^T (G(\mathbf{X}, \mathbf{X}) + \sigma^2 I)^{-1}$. Since GPR variance $\Sigma(\mathbf{x}_*, D)$ is the posterior variance, it is smaller than the prior variance; i.e., $\Sigma(\mathbf{x}_*, D) < g(\mathbf{x}_*, \mathbf{x}_*) + \sigma^2$.

If output $\mathbf{z} \in \mathbb{R}^m$ in (4.2) is multi-dimensional, then GPR is conducted for each output element of $\mathbf{z}$. Let $\mu(\mathbf{x}_*, D(i))$ and $\Sigma(\mathbf{x}_*, D(i))$ denote the GP for the i^{th} element of $\mathbf{z}$ where $D(i) = \langle \mathbf{X}, \mathbf{Z}(i) \rangle$ and $\mathbf{Z}(i) = [\mathbf{z}_1(i), \cdots, \mathbf{z}_N(i)]^T$. Then, we define the Gaussian process function GPR as below

$$[\bar{\mu}(\mathbf{x}_*, D), \bar{\Sigma}(\mathbf{x}_*, D)] = GPR(\mathbf{x}_*, D) \tag{4.5}$$

where mean function $\bar{\mu}(\mathbf{x}_*, D) = [\mu(\mathbf{x}_*, D(1)), \cdots, \mu(\mathbf{x}_*, D(n))]^T$ and variance function $\bar{\Sigma}(\mathbf{x}_*, D) = \text{diag}(\Sigma(\mathbf{x}_*, D(1)), \cdots, \Sigma(\mathbf{x}_*, D(n)))$ denote the approximation of function f and its covariance, respectively.

This chapter utilizes Gaussian kernel, where it is more flexible than other kernels because the RKHS induced by a Gaussian kernel consists of analytic functions whose n^{th} derivative may be non-zero. As shown in (4.3), a different kernel determines different weight shapes. In Gaussian kernel, (i, i') element of G is described by

$$G_{ij} = g(\mathbf{x}_i, \mathbf{x}_j) = \sigma_h^2 e^{-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^T W (\mathbf{x}_i - \mathbf{x}_j)}. \tag{4.6}$$

Diagonal matrix W represents the length scale of each input, and σ_h^2 is a variance. Informally speaking, the length scale $\frac{1}{W_{ll}}$ is a required distance in input space to decouple the output correlation of two inputs. A set of parameters

$$\theta = [W, \sigma_h, \sigma] \tag{4.7}$$

is called hyper-parameters and they show the user's interpretation of the regression function. They may be chosen by maximizing the log-likelihood of the training output so that the choice of hyper-parameters is optimal in some sense:

$$\theta_{\max} = \text{argmax}_\theta(\log(p(\mathbf{Z}|\mathbf{X}, \theta)))$$

as in [2] and [121] where

$$\log(p(\mathbf{Z}|\mathbf{X}, \theta)) = -\frac{1}{2}\mathbf{Z}^T(G(\mathbf{X}, \mathbf{X}) + \sigma^2 I)^{-1}\mathbf{Z} - \frac{1}{2}\log|G(\mathbf{X}, \mathbf{X}) + \sigma^2 I| - \frac{1}{2}N\log 2\pi.$$

In the above equation, the first term is used for data fitting by measurement $\mathbf{Z}$. The second term denotes the complexity penalty. The last term is a normalization constant. Numerical optimization could be performed to find $\theta_{\max}$ such as conjugate gradient descent but there may exist multiple local maxima. To solve the optimization problem, each partial derivative with respect to i^{th} hyper-parameter can be found by $\frac{\partial \log(p(\mathbf{Z}|\mathbf{X}, \theta))}{\partial \theta_i} = \frac{1}{2}tr(\bar{G}^{-1}(\mathbf{X}, \mathbf{X})\mathbf{Z}(\bar{G}^{-1}(\mathbf{X}, \mathbf{X})\mathbf{Z})^T \frac{\partial \bar{G}(\mathbf{X}, \mathbf{X})}{\partial \theta_i})$ where

$$\bar{G}(\mathbf{X}, \mathbf{X}) = G(\mathbf{X}, \mathbf{X}) + \sigma^2 I, \quad \frac{\partial \bar{G}(\mathbf{X}, \mathbf{X})}{\partial \sigma_h} = \frac{2}{\sigma_h} G(\mathbf{X}, \mathbf{X})$$

$$\frac{\partial \bar{G}_{ij}}{\partial W_{ll}} = -\frac{1}{2}(\mathbf{x}_i(l) - \mathbf{x}_j(l))^2 G_{ij} \qquad \frac{\partial \bar{G}(\mathbf{X},\mathbf{X})}{\partial \sigma} = 2\sigma I.$$

Jacobian matrix. Jacobian matrix of GPR is defined by a partial derivative of the mean function with respect to the test input:

$$\frac{\partial \bar{\mu}(\mathbf{x}_*, D)}{\partial \mathbf{x}_*} = \frac{\partial g_*^T (G(\mathbf{X},\mathbf{X}) + \sigma^2 I)^{-1}\mathbf{Z}}{\partial \mathbf{x}_*} = \frac{\partial g_*^T}{\partial \mathbf{x}_*}(G(\mathbf{X},\mathbf{X}) + \sigma^2 I)^{-1}\mathbf{Z}.$$

The Jacobian matrix of GPR will be used to linearize the GPR mean function, and is an approximation of the Jacobian matrix $\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}$ of the regression function.

4.4 Attack-resilient Gaussian process regression

We, in this section, derive a data-driven attack-resilient estimation algorithm to address the problem described in Section 4.2. In Section 4.4.1, we discuss preliminary steps. Then, a description of training data set is given in Section 4.4.2. Section 4.4.3 presents the solution, the Attack-resilient Gaussian Process Regression (ArGPR) algorithm. The proposed algorithm is derived in detail in Section 4.4.4.

4.4.1 Output decomposition, mode, and system transformation

This section discusses two inherent difficulties and tricks to deal with them. First of all, the defender is unaware of the sensor attack locations; i.e., it is unknown which set of sensors is free of attacks. Thus, the defender needs to consider all possible combinations of sensor attack locations. We formulate each possible combination as a hypothetical mode. Let $\mathcal{J}$ denote the set of hypothetical modes, and each mode $j \in \mathcal{J}$ assumes that a particular subset of sensors may be corrupted by sensor attacks, and the others are free of sensor attacks. We denote $y_{1,k}^j$ the outputs of the corrupted sensors and $y_{2,k}^j$ the outputs of the clean sensors. Given mode j, output y_k in (4.1) is decomposed into

$$\begin{aligned} y_{1,k}^j &= C_{1,k}^j x_k + d_{1,k}^j + v_{1,k}^j \\ y_{2,k}^j &= C_{2,k}^j x_k + v_{2,k}^j. \end{aligned} \tag{4.8}$$

Let $s \leq m$ be the number of sensors, then the number of modes is the permutation of the number of sensors $|\mathcal{J}| = 2^p$, where each mode is associated with the corresponding output model (4.8).

Second, state estimation errors and function approximation errors are dependent. To break the interdependency, we let actuator attack vector estimate compensate the actuator attack and the errors of the function approximation. Then, function approximation errors no longer induce errors in state estimation. In particular, we rewrite system (4.1) as follows:

$$x_k = f(x_{k-1}, u_{k-1}) + d'_{2,k-1} + w_{k-1} \tag{4.9}$$

where $d'_{2,k-1} = f(x_{k-1}, u_{k-1}+d_{a,k-1}) - f(x_{k-1}, u_{k-1})$. Given function approximation $\hat{f}_k(\cdot) = \bar{\mu}(\cdot, \hat{D}_k)$ in (4.5) and state estimate $\hat{x}_{k-1}$, system model (4.8) and (4.9) becomes

$$\begin{aligned} x_k &= \hat{f}_k([x_{k-1}^T, u_{k-1}^T]^T) + d_{2,k-1} + w_{k-1} \\ y_{1,k}^j &= C_{1,k}^j x_k + d_{1,k}^j + v_{1,k}^j \\ y_{2,k}^j &= C_{2,k}^j x_k + v_{2,k}^j \end{aligned} \tag{4.10}$$

Our estimation algorithm will utilize linearization to track covariance matrices. Linearization of system (4.10) around the estimates becomes

$$\begin{aligned} x_k &= A_{k-1} x_{k-1} + B_{k-1} u_{k-1} + d_{2,k-1} + w_{k-1} \\ y_{1,k}^j &= C_{1,k}^j x_k + d_{1,k}^j + v_{1,k}^j \\ y_{2,k}^j &= C_{2,k}^j x_k + v_{2,k}^j \end{aligned}$$

where

$$\begin{bmatrix} A_{k-1} \\ B_{k-1} \end{bmatrix} = \frac{\partial \hat{f}([\hat{x}_{k-1}^T, u_{k-1}^T]^T)}{\partial [\hat{x}_{k-1}^T, u_{k-1}^T]^T}$$

4.4.2 Training data set

To regress function f, it is required to know input-output observations according to Section 4.3.2. Let us define $x_k^+ \triangleq f(x_k, u_k) + w_k$. The desired training data set

available at time k is given by:

$$D_k \triangleq \langle \mathbf{X}_k, \mathbf{X}_k^+ \rangle \tag{4.11}$$

where

$$\mathbf{X}_k = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_{N(k)} \end{bmatrix} \triangleq \begin{bmatrix} x_0 & \cdots & x_{k-2} \\ u_0 & \cdots & u_{k-2} \end{bmatrix},$$
$$\mathbf{X}_k^+ = \begin{bmatrix} \mathbf{x}_1^+ & \cdots & \mathbf{x}_{N(k)}^+ \end{bmatrix} \triangleq \begin{bmatrix} x_0^+ & \cdots & x_{k-2}^+ \end{bmatrix}$$

where $N(k)$ is the number of input-output pairs in the training data set. However, unlike Section 4.3.2, x_k and x_k^+ in (4.11) are unavailable. Instead, we will use their estimates to perform function regression. The available training data set $\hat{D}_k$ is given by: Since we will find estimates $\hat{x}_k$ and $\hat{d}_{2,k}$ of the current state x_k and attack vector $d_{2,k}$ at each k, we are able to construct a collection of input-output (estimation) data, where the previous state estimate $[\hat{x}_{k-1}^T, u_{k-1}^T]^T$ be input and the next state estimate minus attack vector estimate $\hat{x}_k - \hat{d}_{2,k}$ be output.

$$\hat{D}_k \triangleq \langle \hat{\mathbf{X}}_k, \hat{\mathbf{X}}_k^+ \rangle \tag{4.12}$$

where

$$\hat{\mathbf{X}}_k = \begin{bmatrix} \hat{\mathbf{x}}_1 & \cdots & \hat{\mathbf{x}}_{N(k)} \end{bmatrix} \triangleq \begin{bmatrix} \hat{x}_0 & \cdots & \hat{x}_{k-2} \\ u_0 & \cdots & u_{k-2} \end{bmatrix},$$

$\hat{\mathbf{X}}_k^+ = [\hat{\mathbf{x}}_1^+, \cdots, \hat{\mathbf{x}}_{N(k)}^+] \triangleq [\hat{x}_1 - \hat{d}_0, \cdots, \hat{x}_{k-1} - \hat{d}_{k-2}]$. Although training data set $\hat{D}_k$ contains estimation errors, we will derive an algorithm as if $\hat{D}_k$ has no estimation errors (certainty equivalence principle [122]). The errors in the training data set will be considered in the analysis section.

4.4.3 Algorithm statement

ArGPR algorithm utilizes Gaussian process regression to approximate unknown dynamic systems, which is then used to estimate the current internal state by unknown input and state estimation technique in Chapter 2. In particular, ArGPR algorithm (Algorithm 6) consists of a bank of ArE algorithm 7 (lines 2-4) for

Algorithm 6 Attack-resilient Gaussian Process Regression (ArGPR)

1: **Input**: $\hat{x}_{k-1}$, P^x_{k-1}, $\hat{D}_k$, ϵ and μ^j_{k-1} for $j \in \mathcal{J}$;
2: **for** $j \in \mathcal{J}$ **do**
3: $\quad [\hat{x}^j_k, \hat{d}^j_{2,k-1}, \hat{d}^j_{1,k}, P^{x^j}_k, P^{d^j_2}_{k-1}, P^{d^j_1}_k, \mathcal{N}^j_k] = ArE(\hat{x}_{k-1},\ P^x_{k-1}, \hat{D}_k)$;
4: **end for**
▷ ***Mode selection***
5: $\bar{\mu}^j_k = \max\{\mathcal{N}^j_k \mu^j_{k-1}, \epsilon\}$;
6: $\mu^j_k = \bar{\mu}^j_k / \sum_{i=1}^{|\mathcal{J}|} \bar{\mu}^i_k$;
7: $\hat{j}_k = \text{argmax}_j(\mu^j_k)$;
▷ ***Training data set update***
8: $\hat{D}_{k+1} = \hat{D}_k \cup \langle [\hat{x}^T_{k-1}, u^T_{k-1}]^T, \hat{x}^{\hat{j}_k}_k - \hat{d}^{\hat{j}_k}_{2,k-1} \rangle$;
9: $\hat{f}_k([x^T, u^T]^T) = \bar{\mu}([x^T, u^T]^T, \hat{D}_k)$;
10: **Return:** $\hat{x}^{\hat{j}_k}_k$, $\hat{d}^{\hat{j}_k}_{2,k-1}$, $\hat{d}^{\hat{j}_k}_{1,k}$, $P^{x^{\hat{j}_k}}_k$, $P^{d^{\hat{j}_k}_2}_{k-1}$, $P^{d^{\hat{j}_k}_1}_k$, $\hat{D}_{k+1}$, and $\hat{f}_k([x^T, u^T]^T)$.

each hypothetical mode as well as a mode estimator (lines 5-7) and a system function estimator (training set updater, line 8). The ArE algorithm can be seen an extension of the extended Kalman filter with two extensions; first, it incorporates attack vector estimation in Chapter 2; second, the unknown system function is replaced by GPR function approximation. The ArE algorithm recursively produces state estimate $\hat{x}^j_k$, attack vector estimates $\hat{d}^j_{2,k-1}$, $\hat{d}_{1,k}$, and prior probability $\mathcal{N}^j_k$ of the associated mode j. The following section presents the algorithm derivation in details.

4.4.4 Derivation of the ArE algorithm

System learning. Given $\hat{D}_k$, we are able to find approximation $\hat{f}_k(\cdot) = \bar{\mu}(\cdot, \hat{D}_k)$ of system function f and covariance $\hat{Q}_k = \bar{\Sigma}(\cdot, \hat{D}_k)$ via $GPR([x^T, u^T]^T, D_k)$ in (4.5). The approximations will be used as if they are the ground truth.

Actuator attack $d_{2,k-1}$ estimation. Assuming that there is no actuator attack, we can predict the current state (line 2) as

$$\hat{x}'_{k|k-1} = \hat{f}_k([\hat{x}^T_{k-1}, u^T_{k-1}]^T).$$

From this estimation, actuator attack $d_{2,k-1}$ can be estimated by $y^j_{2,k}$ in (4.8) as

Algorithm 7 Attack-resilient Estimation (ArE)

1: **Input**: $\hat{x}_{k-1}$, P^x_{k-1}, D_k;
▷ ***Actuator attack*** $d^j_{2,k-1}$ ***estimation***
2: $[\hat{x}'_{k|k-1}, \hat{Q}_{k-1}] = GPR([\hat{x}^T_{k-1}, u^T_{k-1}]^T, D_k)$;
3: $[A^T_{k-1}, B^T_{k-1}]^T = \frac{\partial \bar{\mu}([\hat{x}^T_{k-1}, u^T_{k-1}]^T, D_k)}{\partial [\hat{x}^T_{k-1}, u^T_{k-1}]^T}$;
4: $\hat{y}'^j_{2,k} = C^j_{2,k}\hat{x}'_{k|k-1}$;
5: $\tilde{R}^j_{2,k} = C^j_{2,k}A_{k-1}P^x_{k-1}A^T_{k-1}(C^j_{2,k})^T + R^j_2 + C^j_{2,k}\hat{Q}_{k-1}(C^j_{2,k})^T$;
6: $M^j_k = ((C^j_{2,k})^T(\tilde{R}^j_{2,k})^{-1}C^j_{2,k})^{-1}(C^j_{2,k})^T(\tilde{R}^j_{2,k})^{-1}$;
7: $\hat{d}^j_{2,k-1} = M^j_k(y^j_{2,k} - \hat{y}'^j_{2,k})$;
8: $P^{d^j_2}_{k-1} = M^j_k C^j_{2,k}A_{k-1}P^x_{k-1}(M^j_k C^j_{2,k}A_{k-1})^T + M^j_k R^j_2 (M^j_k)^T + M^j_k C^j_{2,k}\hat{Q}_{k-1}(M^j_k C^j_{2,k})^T$;
▷ ***State prediction***
9: $\hat{x}^j_{k|k-1} = \hat{f}_k([\hat{x}^T_{k-1}, u^T_{k-1}]^T) + \hat{d}^j_{2,k-1}$;
10: $\bar{Q}^j_{k-1} = (I - M^j_k C^j_{2,k})\hat{Q}_{k-1}(I - M^j_k C^j_{2,k})^T + M^j_k R^j_2 (M^j_k)^T$;
11: $\bar{A}^j_{k-1} = (I - M^j_k C^j_{2,k})A_{k-1}$;
12: $P^{x^j}_{k|k-1} = \bar{A}^j_{k-1}P^x_{k-1}(\bar{A}^j_{k-1})^T + \bar{Q}^j_{k-1}$;
▷ ***State estimation***
13: $L^j_k = (P^{x^j}_{k|k-1}(C^j_{2,k})^T - M^j_k R^j_2)\ (C^j_{2,k}P^{x^j}_{k|k-1}(C^j_{2,k})^T + R^j_2 - C^j_{2,k}M^j_k R^j_2 - R^j_2(M^j_k)^T(C^j_{2,k})^T)^{-1}$;
14: $\hat{x}^j_k = \hat{x}^j_{k|k-1} + L^j_k(y^j_{2,k} - C^j_{2,k}\hat{x}_{k|k-1})$;
15: $P^{x^j}_k = (I - L^j_k C^j_{2,k})P^{x^j}_{k|k-1}(I - L^j_k C^j_{2,k})^T + L^j_k R^j_2 (L^j_k)^T + (I - L^j_k C^j_{2,k})M^j_k R^j_2 (L^j_k)^T + L^j_k R^j_2 (M^j_k)^T (I - L^j_k C^j_{2,k})^T$;
▷ ***Sensor attack*** $d^j_{1,k}$ ***estimation***
16: $\hat{d}^j_{1,k} = y^j_{1,k} - C^j_{1,k}\hat{x}^j_k$;
17: $P^{d^j_1}_k = C^j_{1,k}P^{x^j}_k(C^j_{1,k})^T + R^j_1$;
▷ ***The priori probability of mode***
18: $\nu^j_k = y^j_{2,k} - \hat{y}^j_{2,k}$;
19: $\bar{P}^j_{k|k-1} = C^j_{2,k}P^{x^j}_{k|k-1}(C^j_{2,k})^T + R^j_2$;
20: $\mathcal{N}^j_k = \frac{\exp(-(\nu^j_k)^T(\bar{P}^j_{k|k-1})^{-1}\nu^j_k/2)}{(2\pi)^{dim(y^j_{2,k})/2}|\bar{P}^j_{k|k-1}|^{1/2}}$;
21: **Return:** $\hat{x}^j_k$, $\hat{d}^j_{2,k-1}$, $\hat{d}^j_{1,k}$, $P^{x^j}_k$, $P^{d^j_2}_{k-1}$, $P^{d^j_1}_k$, η^j_k.

follows (line 7):

$$\hat{d}^j_{2,k-1} = M^j_k(y^j_{2,k} - C^j_{2,k}\hat{x}'_{k|k-1})$$

$$\simeq M_k^j(C_{2,k}^j A_{k-1}\tilde{x}_{k-1} + C_{2,k}^j d_{2,k-1} + C_{2,k-1}^j w_{k-1} + v_{2,k}^j)$$

where function $\hat{f}_k([\hat{x}_{k-1}^T, u_{k-1}^T]^T)$ is linearized. Assuming that $\mathbb{E}[\tilde{x}_{k-1}] = 0$, we choose gain matrix M_k^j by the Gauss Markov theorem (Theorem 2.8.1) as follows (line 6):

$$M_k^j = ((C_{k,2}^j)^T(\tilde{R}_{2,k}^j)^{-1}C_{2,k}^j)^{-1}(C_{2,k}^j)^T(\tilde{R}_{2,k}^j)^{-1}$$

where $\tilde{R}_{2,k}^j \triangleq C_{2,k}^j A_{k-1}P_{k-1}^x A_{k-1}^T(C_{2,k}^j)^T + C_{2,k}^j\hat{Q}_{k-1}(C_{2,k}^j)^T + R_{2,k}^j$. Estimation error of $d_{2,k-1}$ is given by

$$\tilde{d}_{2,k-1}^j = -M_k^j(C_{2,k}^j A_{k-1}\tilde{x}_{k-1} + C_{2,k}^j w_{k-1} + v_{2,k}^j) \tag{4.13}$$

where the property $M_k^j C_{2,k}^j = I$ is used. Error covariance matrix is found by $P_{k-1}^{d_2^j} = M_k^j C_{2,k}^j A_{k-1}P_{k-1}^x(M_k^j C_{2,k}^j A_{k-1})^T + M_k^j C_{2,k}^j\hat{Q}_{k-1}(M_k^j C_{2,k}^j)^T + M_k^j R_{2,k}^j(M_k^j)^T$ (line 8).

State prediction. Generate state prediction $\hat{x}_{k|k-1}^j$ by simulating system (4.1) as follows (line 9):

$$\hat{x}_{k|k-1}^j = \hat{f}_k([\hat{x}_{k-1}^T, u_{k-1}^T]^T) + \hat{d}_{2,k-1}^j \tag{4.14}$$

The state estimation error becomes

$$\tilde{x}_{k|k-1}^j \simeq A_{k-1}\tilde{x}_{k-1} + \tilde{d}_{2,k-1}^j + w_{k-1} \tag{4.15}$$

where function $\hat{f}_k$ is linearized. Substitution (4.13) into (4.15) leads to (line 12)

$$P_{k|k-1}^{x^j} = \bar{A}_{k-1}^j P_{k-1}^x(\bar{A}_{k-1}^j)^T + \bar{Q}_{k-1}^j. \tag{4.16}$$

where $\bar{Q}_{k-1}^j \triangleq (I - M_k^j C_{2,k}^j)\hat{Q}_{k-1}(I - M_k^j C_{2,k}^j)^T + M_k^j R_{2,k}^j(M_k^j)^T$, and $\bar{A}_{k-1}^j \triangleq (I - M_k^j C_{2,k}^j)A_{k-1}$.

State estimation. Update state prediction $\hat{x}_{k|k-1}^j$ as (line 14)

$$\hat{x}_k^j = \hat{x}_{k|k-1}^j + L_k^j(y_{2,k}^j - C_{2,k}^j\hat{x}_{k|k-1}^j). \tag{4.17}$$

By substituting $y_{2,k}^j$ in (4.8) into (4.17), the state estimation error becomes

$$\tilde{x}_k^j \simeq (I - L_k^j C_{2,k}^j)\tilde{x}_{k|k-1}^j - L_k^j v_{2,k}^j. \tag{4.18}$$

Its error covariance matrix is (line 15)

$$\begin{aligned} P_k^{x^j} &= (I - L_k^j C_{2,k}^j) P_{k|k-1}^{x^j} (I - L_k^j C_{2,k}^j)^T + L_k^j R_{2,k}^j (L_k^j)^T + (I - L_k^j C_{2,k}^j) M_k^j R_{2,k}^j (L_k^j)^T \\ &\quad + L_k^j R_{2,k}^j (M_k^j)^T (I - L_k^j C_{2,k}^j)^T \end{aligned} \tag{4.19}$$

Minimizing error covariance $\mathrm{tr}(P_k^{x^j})$ with decision variable L_k^j is an unconstrained optimization problem. We can find the minimizer by taking derivative of $\mathrm{tr}(P_k^{x^j})$ and setting it equal to zero

$$\begin{aligned} \frac{\partial \,\mathrm{tr}(P_k^{x^j})}{\partial L_k^j} &= 2((C_{2,k}^j P_{k|k-1}^{x^j} (C_{2,k}^j)^T - R_{2,k}^j (M_k^j)^T (C_{2,k}^j)^T \\ &\quad - C_{2,k}^j M_k^j R_{2,k}^j + R_{2,k}^j)(L_k^j)^T + R_{2,k}^j (M_k^j)^T - C_{2,k}^j P_{k|k-1}^{x^j}). \end{aligned}$$

The solution is $L_k^j = (P_{k|k-1}^{x^j}(C_{2,k}^j)^T - M_k^j R_{2,k}^j)(R_{2,k}^j + C_{2,k}^j P_{k|k-1}^{x^j}(C_{2,k}^j)^T - C_{2,k}^j M_k^j R_{2,k}^j - R_{2,k}^j (M_k^j)^T (C_{2,k}^j)^T)^{-1}$ (line 13).

Sensor attack $d_{1,k}^j$ estimation. Given $\hat{x}_k^j$, and the assumption that $\mathbb{E}[\tilde{x}_k^j] = 0$, sensor attack $d_{1,k}^j$ can be estimated by $y_{1,k}^j$ in (4.8) (line 16):

$$\hat{d}_{1,k}^j = y_{1,k}^j - C_{1,k}^j \hat{x}_k = C_{1,k}^j \tilde{x}_k^j + d_{1,k}^j + v_{1,k}^j. \tag{4.20}$$

Estimation error $\tilde{d}_{1,k}^j$ is obtained by

$$\tilde{d}_{1,k}^j = -(C_{1,k}^j \tilde{x}_k^j + v_{1,k}^j) \tag{4.21}$$

with error covariance matrix $P_k^{d_1^j} = C_{1,k}^j P_k^{x^j} (C_{1,k}^j)^T + R_{1,k}^j$ (line 17).

The probability of mode. It is natural that the predicted output must be matched with the measured output if the mode j is the true mode. For $\forall j \in \mathcal{J}$, we quantify the difference between the predicted output and the measured output as follows (line 18)

$$\nu_k^j = y_{2,k}^j - C_{2,k}^j \hat{x}_{k|k-1}^j.$$

The output error ν_k^j is a multivariate Gaussian random variable. The likelihood function is given by (line 20)

$$\mathcal{N}_k^j \triangleq p(y_k|j = \text{true}) = \mathcal{N}(\nu_k^j; 0, \bar{P}_{k|k-1}^j) = \frac{\exp(-(\nu_k^j)^T(\bar{P}_{k|k-1}^j)^{-1}\nu_k^j/2)}{(2\pi)^{dim(y_{2,k}^j)/2}|\bar{P}_{k|k-1}^j|^{\frac{1}{2}}}$$

where $\bar{P}_{k|k-1}^j = C_{2,k}^j P_{k|k-1}^j (C_{2,k}^j)^T + R_{2,k}^j$ is the error covariance matrix of ν_k^j (line 19).

Mode selection (ArGPR). By the Bayes' theorem, the posterior probability is

$$\begin{aligned}\mu_k^j &\triangleq p(j = \text{true}|y_k, \cdots, y_0) = \frac{p(y_k|j = \text{true})p(j = \text{true}|y_{k-1}, \cdots, y_0)}{\sum_{i=1}^{|\mathcal{J}|} p(y_k|j = \text{true})p(j = \text{true}|y_{k-1}, \cdots, y_0)} \\ &= \frac{\mathcal{N}_k^j \mu_{k-1}^j}{\sum_{i=1}^{|\mathcal{J}|} \mathcal{N}_k^j \mu_{k-1}^j}.\end{aligned}$$

However, such update might allow that some μ_k^j converge to zero. To prevent this, we modify the posterior probability update to (lines 5-6) $\mu_k^j = \frac{\bar{\mu}_k^j}{\sum_{i=1}^{|\mathcal{J}|} \bar{\mu}_k^i}$, where $\bar{\mu}_k^j = \max\{\mathcal{N}_k^j \mu_{k-1}^j, \epsilon\}$ and $0 < \epsilon < \frac{1}{|\mathcal{J}|}$ is a pre-selected small constant preventing the vanishment of the mode probability. The most likely mode is chosen as the current mode $\hat{j}_k = \text{argmax}_j(\mu_k^j)$ (line 7).

Training data set update (ArGPR). Lastly, we construct a new training data pair and add it to the training data set (line 8)

$$D_{k+1} = D_k \cup \langle [\hat{x}_{k-1}^T, u_{k-1}^T]^T, \hat{x}_k^{\hat{j}} - \hat{d}_{2,k-1}^{\hat{j}_k} \rangle.$$

Then, the algorithm returns estimates of system function and output function from the updated training data set as follows

$$\hat{f}_k([x^T, u^T]^T) = \bar{\mu}([x^T, u^T]^T, D_k). \tag{4.22}$$

4.5 Analysis

We study stability of estimation errors and function approximation errors, given that $|\mathcal{J}| = 1$ and $\mathcal{J}$ is known. Since $|\mathcal{J}| = 1$, we drop index j in this section.

Section 4.5.1 studies stability of state estimation errors, where the analysis is independent of function approximation errors because the function approximation errors do not influence state estimation errors. Section 4.5.2 presents an analysis of function approximation errors. In particular, average case learning of GPR is discussed.

4.5.1 Estimation error analysis

We consider linearization error ϕ_k defined by

$$\hat{f}_k([x_{k-1}^T, u_{k-1}^T]^T) - \hat{f}_k([\hat{x}_{k-1}^T, u_{k-1}^T]^T) = A_{k-1}\tilde{x}_{k-1} + \phi_{k-1}(\hat{x}_{k-1}, x_{k-1}, u_{k-1}).$$

The following set of assumptions is needed to ensure the stability of the ArE algorithm.

Assumption 4.5.1 *Matrix $C_{2,k}$ has full column rank.*

Assumption 4.5.2 *There exist $\bar{a}$, $\bar{c}_1$, $\bar{c}_2$, $\underline{q}$, $\underline{r}_2 > 0$ such that the following holds for $k \geq 0$:*

$$\|A_k\| \leq \bar{a},\ \|C_{1,k}\| \leq \bar{c}_1,\ \underline{c}_2 \leq \|C_{2,k}\| \leq \bar{c}_2,\ \underline{q} \leq \hat{Q}_k,\ \underline{r}_2 I \leq R_{2,k}.$$

Assumption 4.5.3 *For any $\epsilon_\phi > 0$, there exists $\delta > 0$ such that*

$$\|\phi_k(\hat{x}_k, x_k, u_k)\| \leq \epsilon_\phi \|x_k - \hat{x}_k\|^2$$

holds for all $\|x_k - \hat{x}_k\| \leq \delta$ and $k \geq 0$.

In Assumption 4.5.2, $\|A_k\| \leq \bar{a}$ holds if $\hat{f}_k$ is Lipschitz. Assumption 4.5.3 holds if $\hat{f}_k$ is Holder continuous (p.136 in [97]) with exponent 2.

Theorem 4.5.1 *(Stability in the presence of modeling uncertainties) Consider the ArGPR algorithm, provided that Assumptions 4.5.1, 4.5.2, and 4.5.3. For any $\gamma \in (0, 1)$, there exists a set of positive constants α_x, α_{d_1}, α_{d_2}, b_x, b_{d_1}, b_{d_2}, c_x, c_{d_1}, c_{d_2}, $\underline{\delta}$, $\bar{q}'$, $\bar{r}_1$, $\bar{r}_2$, and $\bar{\epsilon}$ such that, if $\hat{Q}_k \leq \bar{q}' I$, $R_{1,k} \leq \bar{r}_1 I$, $R_{2,k} \leq \bar{r}_2 I$, and $\epsilon \leq \bar{\epsilon}$, then the following properties hold:*

$$P(\|\tilde{x}_k\| < \alpha_x e^{-b_x k}\|\tilde{x}_0\| + c_x) \geq 1 - \gamma,$$

$$P(\|\tilde{d}_{1,k}\| < \alpha_{d_1} e^{-b_{d_1} k}\|\tilde{x}_0\| + c_{d_1}) \geq 1-\gamma,$$
$$P(\|\tilde{d}_{2,k}\| < \alpha_{d_2} e^{-b_{d_2} k}\|\tilde{x}_0\| + c_{d_2}) \geq 1-\gamma$$

for all $\|\tilde{x}_0\| \leq \underline{\delta}$ *and* $k \geq 0$.

PROOF. The proof is similar to that of Theorem 2.5.1 in Chapter 2. We omit its details. ■

Theorem 4.5.1 shows that estimation errors are stable in probability. Since $d_{2,k-1} = d'_{2,k-1} + \tilde{f}_k(x_{k-1}, u_{k-1})$, stability of $\tilde{d}_{2,k}$ implies that $d_{2,k}$ estimate both actuator attacks and function approximation errors; i.e., $\hat{d}_{2,k}$ compensates those uncertainties. If the function approximation errors are sufficiently small, $\hat{d}_{2,k}$ represents actuator attacks. The following section studies how the function approximation error behaves.

4.5.2 Average case learning of GPR

We, in this section, analyze how average case learning of GPR behaves. In particular, the error

$$f(x_{k-1}, u_{k-1}) - \hat{f}_k([\hat{x}_{k-1}^T, u_{k-1}^T]^T)$$

is the point of interest. The above error becomes $\tilde{f}_k(x_{k-1}, u_{k-1})$ if $x_{k-1} = \hat{x}_{k-1}$.

Training data set. Let us define the errors in the training data set as follows: $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \hat{\mathbf{x}}_i$, $\tilde{\mathbf{X}}_k = \mathbf{X}_k - \hat{\mathbf{X}}_k$, $\tilde{\mathbf{x}}_i^+ = \mathbf{x}_i^+ - \hat{\mathbf{x}}_i^+$, and $\tilde{\mathbf{X}}_k^+ = \mathbf{X}_k^+ - \hat{\mathbf{X}}_k^+$. Using the above notations, training data set $\hat{D}_k$ in (4.12) becomes

$$\begin{aligned}
\hat{D}_k &\triangleq \langle \hat{\mathbf{X}}_k, \hat{\mathbf{X}}_k^+ \rangle \\
&= \langle \hat{\mathbf{X}}_k, \mathbf{f}(\hat{\mathbf{X}}_k) + \mathbf{w}_k - \mathbf{f}(\hat{\mathbf{X}}_k) + \mathbf{f}(\mathbf{X}_k) - \tilde{\mathbf{X}}_k^+ \rangle \\
&= \langle \hat{\mathbf{X}}_k, \hat{\mathbf{Z}}_k - \mathbf{f}(\hat{\mathbf{X}}_k) + \mathbf{f}(\mathbf{X}_k) - \tilde{\mathbf{X}}_k^+ \rangle \\
&= \langle \hat{\mathbf{X}}_k, \hat{\mathbf{Z}}_k + \tilde{\mathbf{Z}}_k \rangle
\end{aligned} \tag{4.23}$$

where $\mathbf{f}(\hat{\mathbf{X}}_k) = [f(\hat{\mathbf{x}}_1), \cdots, f(\hat{\mathbf{x}}_{N(k)})]$ and $\hat{\mathbf{Z}}_k = \mathbf{f}(\hat{\mathbf{X}}_k) + \mathbf{w}_k$. In the analysis, we use $\hat{\mathbf{X}}_k$ and $\hat{\mathbf{Z}}_k$ as the input and output to learn function f as the classic GPR in Section 4.3.2. Correspondingly, we consider $\tilde{\mathbf{Z}}_k = [\tilde{\mathbf{z}}_1, \cdots, \tilde{\mathbf{z}}_{N(k)}] = -\mathbf{f}(\hat{\mathbf{X}}_k) + \mathbf{f}(\mathbf{X}_k) - \tilde{\mathbf{X}}_k^+$ be the output errors.

Analysis in Reproducing Kernel Hilbert Space. Most widely used kernels, including Gaussian kernel (4.6), satisfy the following assumption.

Assumption 4.5.4 *Kernel g is continuous symmetric and positive definite. Kernel is time-invariant.*

Hyper-parameter θ in (4.7) is chosen time-invariant to satisfy Assumption 4.5.4. Under Assumption 4.5.4, there exists a unique Reproducing Kernel Hilbert Space $\mathcal{H}$ (RKHS) by the Moore-Aronszajn theorem (Theorem 6.1 in [2]).

Now consider the minimization of functional

$$J_k[\hat{f}] = \frac{1}{2}\|\hat{f}\|_{\mathcal{H}}^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{N(k)}(\hat{\mathbf{x}}_i^+ - \hat{f}(\hat{\mathbf{x}}_i))^2 \tag{4.24}$$

where σ is the variance of w in (4.2). The second term works for data fitting and the first term smooths the solution, called regularizer. According to Section 6.2.2 in [2], the minimizer of functional (4.24) is the GPR mean function (4.3). In particular, the minimizer of the above functional is in the form of

$$\hat{f}(\mathbf{x}_*) = \sum_{i=1}^{N(k)} \alpha_i g(\mathbf{x}_*, \hat{\mathbf{x}}_i) \tag{4.25}$$

by the representer theorem [123]. Utilizing (4.25), functional (4.24) becomes

$$J_k[\alpha] = \frac{1}{2}\alpha^T G(\hat{\mathbf{X}}_k, \hat{\mathbf{X}}_k)\alpha + \frac{1}{2\sigma^2}\|\hat{\mathbf{X}}_k^+ - G(\hat{\mathbf{X}}_k, \hat{\mathbf{X}}_k)\alpha\|^2$$

where $\alpha = [\alpha_1, \cdots, \alpha_{N(k)}]^T$. By taking its derivative with respect to vector α and setting it equal to zero, we can obtain the solution $\alpha = (G(\hat{\mathbf{X}}_k, \hat{\mathbf{X}}_k) + \sigma^2 I)^{-1}\hat{\mathbf{X}}_k^+$. The complete solution $\hat{f}(\mathbf{x}_*) = g_*^T(\mathbf{x}_*)(G(\hat{\mathbf{X}}_k, \hat{\mathbf{X}}_k) + \sigma^2 I)^{-1}\hat{\mathbf{X}}_k^+$ is identical to the GPR mean function in (4.5). Motived by this property, let us define

$$\tilde{f}_{k|J_k}(x_{k-1}, u_{k-1}, \hat{x}_{k-1}) \triangleq f(x_{k-1}, u_{k-1}) - \hat{f}_{k|J_k}(\hat{x}_{k-1})$$

where $\hat{f}_{k|J_k} = \text{argmin}_{\hat{f}} J_k[\hat{f}]$. Note that $\tilde{f}_{k|J_k} = \tilde{f}_k$ and $\hat{f}_{k|J_k} = \hat{f}_k$, provided that $\hat{D}_k$ is known and $\hat{x}_{k-1} = x_{k-1}$ holds. We will analyze average case GPR learning $\hat{f}_{k|\mathbb{E}[J_k]}$ under the following assumptions.

Assumption 4.5.5 *The regression function f is in RKHS $\mathcal{H}$.*

Since $g \in \mathcal{H}$, any linear combination of g is in RKHS $\mathcal{H}$. Thus, function f is in RKHS $\mathcal{H}$ if and only if there exist a set of $\mathbf{x}_i \in \mathbb{R}^{n+a}$, and $\beta_i \in \mathbb{R}$ such that

$$f(\mathbf{x}) = \sum_{i=1}^{\infty} \beta_i g(\mathbf{x}_i, \mathbf{x}). \tag{4.26}$$

Comparing (4.26) with (4.25), Assumption 4.5.5 implies that the chosen kernel can perform sufficiently well to approximate the regression function.

Assumption 4.5.6 *Input $\hat{\boldsymbol{X}}_k$ and output $\hat{\boldsymbol{Z}}_k$ in training data are sampled from probability distributions with corresponding probability measure $\mu(\hat{\boldsymbol{x}}, \hat{\boldsymbol{z}})$. State estimation errors $\tilde{\boldsymbol{X}}_k$, $\tilde{\boldsymbol{X}}_k^+$ are independent of $\boldsymbol{X}_k$ and $\hat{\boldsymbol{Z}}_k$, respectively.*

Under Assumption 4.5.4, according to Mercer's theorem (Theorem 4.2 in [2]), there is a set of orthonormal eigenfunctions $\{\phi_j\}$ and nonnegative eigenvalues $\{\lambda_j\}$ corresponding to the kernel such that $g(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{\infty} \lambda_j \phi_j(\mathbf{x}) \phi_j(\mathbf{x}')$ where $\sum_{j=1}^{\infty} \lambda_j < \infty$. Under Assumption 4.5.5, there exists a set of constants $c_j \in \mathbb{R}$ such that $f(\mathbf{x}_*) = \sum_{j=1}^{\infty} c_j \phi_j(\mathbf{x}_*)$. There are infinitely many set of orthonomal eigenfunctions. Of them, we choose one such that Assumption 4.5.7 holds; e.g., $\phi_j(\mathbf{x}) = e^{\sqrt{-1}^2 j\mathbf{x}}$ (Fourier transform).

Assumption 4.5.7 *Eigenfunctions satisfy $\phi_j(\boldsymbol{x} + \tilde{\boldsymbol{x}}) = \phi_j(\boldsymbol{x})\phi_j(\tilde{\boldsymbol{x}})$.*

Theorem 4.5.2 *Under Assumptions 4.5.4, 4.5.5, 4.5.6 and 4.5.7, it holds that*

$$\begin{aligned}\hat{f}_{k|\mathbb{E}[J_k]}(\boldsymbol{x}_*) = \sum_{j=1}^{\infty} \Big[\frac{\lambda_j}{\lambda_j + \sigma^2/N(k)} \frac{1}{N(k)} \sum_{i=1}^{N(k)} (c_j \phi_j(-\tilde{\boldsymbol{x}}_i) \\ + \tilde{\boldsymbol{x}}_i \int \phi_j(\hat{\boldsymbol{x}}) d\mu(\hat{\boldsymbol{x}})) \phi_j(\boldsymbol{x}_*) \Big].\end{aligned}$$

PROOF. Consider functional $\mathbb{E}[J_k]$ and let $\hat{f}_k = \hat{f}_{k|\mathbb{E}[J_k]}$ in this proof for notational simplicity. Probability measure $\mu(\hat{\mathbf{x}}, \hat{\mathbf{z}})$ is accosiated with random distributions of $\hat{\mathbf{x}}$, and $\hat{\mathbf{z}}$.

Since $\hat{\mathbf{x}}_i^+ - \hat{f}(\hat{\mathbf{x}}_i) = \hat{\mathbf{z}}_i - f(\hat{\mathbf{x}}_i) + f(\mathbf{x}_i) - \tilde{\mathbf{x}}_i^+ - \hat{f}(\hat{\mathbf{x}}_i)$, we have

$$\begin{aligned}\mathbb{E}[(\hat{\mathbf{x}}_i^+ - \hat{f}(\hat{\mathbf{x}}_i))^2] &= \int (\mathbf{z}_i - \tilde{\mathbf{x}}_i^+ - \hat{f}(\hat{\mathbf{x}}_i))^2 d\mu(\hat{\mathbf{x}}_i, \hat{\mathbf{z}}_i) \\ &= \int (\hat{\mathbf{z}}_i - f(\hat{\mathbf{x}}_i))^2 d\mu(\hat{\mathbf{x}}_i, \hat{\mathbf{z}}_i) + \int (f(\mathbf{x}_i) - \hat{f}(\hat{\mathbf{x}}_i))^2 d\mu(\hat{\mathbf{x}}_i, \hat{\mathbf{z}}_i)\end{aligned}$$

$$+\int(\tilde{\mathbf{x}}_i^+)^2 d\mu(\hat{\mathbf{x}}_i,\hat{\mathbf{z}}_i)+\int 2(\hat{\mathbf{z}}_i-f(\hat{\mathbf{x}}_i))(f(\mathbf{x}_i)-\hat{f}(\hat{\mathbf{x}}_i))d\mu(\hat{\mathbf{x}}_i,\hat{\mathbf{z}}_i)$$
$$-\int 2\tilde{\mathbf{x}}_i^+(\hat{\mathbf{z}}_i-f(\hat{\mathbf{x}}_i)+f(\mathbf{x}_i)-\hat{f}(\hat{\mathbf{x}}_i))d\mu(\hat{\mathbf{x}}_i,\hat{\mathbf{z}}_i).$$

Since $\int d\mu(\hat{\mathbf{x}}_i,\hat{\mathbf{z}}_i) = \int\int \mu(\hat{\mathbf{z}}_i|\hat{\mathbf{x}}_i)\mu(\hat{\mathbf{x}}_i)$, $\int(\hat{\mathbf{z}}_i - f(\hat{\mathbf{x}}_i))^2 d\mu(\hat{\mathbf{z}}_i|\hat{\mathbf{x}}_i) = \sigma^2$ and $\int \hat{\mathbf{z}}_i - f(\hat{\mathbf{x}}_i)d\mu(\hat{\mathbf{z}}_i|\hat{\mathbf{x}}_i) = 0$, the above equation becomes

$$\mathbb{E}[(\hat{\mathbf{x}}_i^+ - \hat{f}(\hat{\mathbf{x}}_i))^2] = \int (f(\mathbf{x}_i)-\hat{f}(\hat{\mathbf{x}}_i))^2 d\mu(\hat{\mathbf{x}}_i) + (\tilde{\mathbf{x}}_i^+)^2 + \int \sigma^2 d\mu(\hat{\mathbf{x}}_i)$$
$$-\tilde{\mathbf{x}}_i^+ \int 2(f(\mathbf{x}_i)-\hat{f}(\hat{\mathbf{x}}_i))d\mu(\hat{\mathbf{x}}_i) \tag{4.27}$$

where $(\tilde{\mathbf{x}}_i^+)^2$, and $\int \sigma^2 d\mu(\hat{\mathbf{x}}_i)$ are independent of $\hat{f}$. Substituting (4.27) without those independent terms into $\mathbb{E}[J[\hat{f}]]$, we have the functional to minimize:

$$\mathbb{E}'[J[\hat{f}]] = \frac{1}{2}\|\hat{f}\|_{\mathcal{H}}^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{N(k)}(\int (f(\mathbf{x}_i)-\hat{f}(\hat{\mathbf{x}}_i))^2 d\mu(\hat{\mathbf{x}}_i)$$
$$-\tilde{\mathbf{x}}_i^+ \int 2(f(\mathbf{x}_i)-\hat{f}(\hat{\mathbf{x}}_i))d\mu(\hat{\mathbf{x}}_i)). \tag{4.28}$$

Since f and $\hat{f}$ are in RKHS, it can be expressed using eigenfunctions

$$f(\mathbf{x}) = \sum_{j=1}^{\infty} c_j\phi_j(\mathbf{x}) = \sum_{j=1}^{\infty} c_j\phi_j(\hat{\mathbf{x}})\phi_j(\tilde{\mathbf{x}}),$$
$$\hat{f}(\hat{\mathbf{x}}) = \sum_{j=1}^{\infty} \hat{c}_j\phi_j(\hat{\mathbf{x}}). \tag{4.29}$$

Note that $\langle f, f'\rangle_{\mathcal{H}} = \sum_{j=1}^{\infty}\frac{c_j c_j'}{\lambda_j}$. Also, the eigenfunctions are orthogonal to each other $\int \phi_i(\hat{\mathbf{x}})\phi_j(\hat{\mathbf{x}})d\mu(\hat{\mathbf{x}}) = \delta_{ij}$ where δ_{ij} is Kronecker delta. By substituting (4.29) into (4.28), we have

$$\mathbb{E}'[J[\hat{c}]] = \frac{1}{2}\sum_{j=1}^{\infty}\frac{\hat{c}_j^2}{\lambda_j} + \frac{1}{2\sigma^2}\sum_{i=1}^{N(k)}(\int(\sum_{j=1}^{\infty}(c_j\phi_j(\tilde{\mathbf{x}}_i)-\hat{c}_j)\phi_j(\hat{\mathbf{x}}))^2 d\mu(\hat{\mathbf{x}})$$
$$-2\tilde{\mathbf{x}}_i^+\int\sum_{j=1}^{\infty}(c_j\phi_j(\tilde{\mathbf{x}}_i)-\hat{c}_j)\phi_j(\hat{\mathbf{x}})d\mu(\hat{\mathbf{x}}))$$
$$= \frac{1}{2}\sum_{j=1}^{\infty}\frac{\hat{c}_j^2}{\lambda_j} + \frac{1}{2\sigma^2}\sum_{i=1}^{N(k)}(\sum_{j=1}^{\infty}(c_j\phi_j(\tilde{\mathbf{x}}_i)-\hat{c}_j)^2 - 2\tilde{\mathbf{x}}_i^+\sum_{j=1}^{\infty}(c_j\phi_j(\tilde{\mathbf{x}}_i)-\hat{c}_j)\int\phi_j(\hat{\mathbf{x}})d\mu(\hat{\mathbf{x}}))$$

where $\hat{c} = [c_1, \cdots, c_{N(k)}]^T$. It is a convex optimization problem with respect to decision variables $\hat{c}$. By taking derivative with respect to $\hat{c}_j$ and setting it equal to zero, we have

$$\frac{\hat{c}_j}{\lambda_j} + \frac{1}{\sigma^2}\sum_{i=1}^{N(k)}(-c_j\phi_j(\tilde{\mathbf{x}}_i) + \hat{c}_j + \tilde{\mathbf{x}}_i^+ \int \phi_j(\hat{\mathbf{x}})d\mu(\hat{\mathbf{x}})) = 0.$$

The solution is

$$\hat{c}_j = \frac{\lambda_j}{\lambda_j + \sigma^2/N(k)}\frac{1}{N(k)}\sum_{i=1}^{N(k)}(c_j\phi_j(\tilde{\mathbf{x}}_i) - \tilde{\mathbf{x}}_i^+ \int \phi_j(\hat{\mathbf{x}})d\mu(\hat{\mathbf{x}})).$$

Plugging the solution into $\hat{f}(\mathbf{x}) = \sum_{i=1}^{\infty}\hat{c}_i\phi_i(\mathbf{x})$, we have the desired result. ■

Interpretation of Theorem 4.5.2. According to Theorem 4.5.2, identification error is described by

$$\begin{aligned}
&\tilde{f}_{k|\mathbb{E}[J_k]}(x_{k-1}, u_{k-1}, \hat{x}_{k-1}) = f(\mathbf{x}_{k-1}) - \hat{f}_{k|\mathbb{E}[J_k]}(\hat{\mathbf{x}}_{k-1}) \\
&= f(\mathbf{x}_{k-1}) - \hat{f}_{k|\mathbb{E}[J_k]}(\mathbf{x}_{k-1}) + \hat{f}_{k|\mathbb{E}[J_k]}(\mathbf{x}_{k-1}) - \hat{f}_{k|\mathbb{E}[J_k]}(\hat{\mathbf{x}}_{k-1}) \\
&= \sum_{j=1}^{\infty}(c_j - \frac{\lambda_j}{\lambda_j + \sigma^2/N(k)}c_j)\phi_j(\mathbf{x}_{k-1}) \\
&+ \sum_{j=1}^{\infty}\frac{\lambda_j}{\lambda_j + \sigma^2/N(k)}(c_j - c_j\frac{1}{N(k)}\sum_{i=1}^{N(k)}\phi_j(\tilde{\mathbf{x}}_i))\phi_j(\mathbf{x}_{k-1}) \\
&- \sum_{j=1}^{\infty}\frac{\lambda_j}{\lambda_j + \sigma^2/N(k)}\int \phi_j(\hat{\mathbf{x}})d\mu(\hat{\mathbf{x}})\frac{1}{N(k)}\sum_{i=1}^{N(k)}\tilde{\mathbf{x}}_i^+\phi_j(\hat{\mathbf{x}}_{k-1}) \\
&+ \sum_{j=1}^{\infty}\frac{\lambda_j}{\lambda_j + \sigma^2/N(k)}\frac{1}{N(k)}\sum_{i=1}^{N(k)}(c_j\phi_j(\tilde{\mathbf{x}}_i) - \tilde{\mathbf{x}}_i^+ \int \phi_j(\hat{\mathbf{x}})d\mu(\hat{\mathbf{x}})) \\
&\times (\phi_j(\mathbf{x}_{k-1}) - \phi_j(\hat{\mathbf{x}}_{k-1})).
\end{aligned}$$

The first term indicates the function identification error provided that $\tilde{\mathbf{X}}_k = 0$ and $\tilde{\mathbf{X}}_k^+ = 0$. This error decreases as $N(k) \to \infty$, where $\lim_{k\to\infty} N(k) = \lim_{k\to\infty} k + N(0) = \infty$.

The second term is the error induced by $\tilde{\mathbf{X}}_k$. Note that $\phi_i(0) = 1$ under Assumption 4.5.7. Thus, this error is zero if $\tilde{\mathbf{X}}_k = 0$. If there is the finite number of indices i such that $\tilde{\mathbf{x}}_i \neq 0$, then $\lim_{N(k)\to\infty} c_j - c_j\frac{1}{N(k)}\sum_{i=1}^{N(k)}\phi_j(\tilde{\mathbf{x}}_i) = 0$; i.e., the finite number of errors are overwhelmed by a number of correct training data.

Table 4.1: Variables and parameters of IEEE 68-bus test system.

System variables			
f	angular frequency	θ	phase angle
P_M	mechanical power	P_{ij}	power flow
P_C	controllable load	P_L	net load
P_{elec}	electrical power output		
System parameters			
D	damping constant	m	angular momentum
t_{ij}	tie-line stiffness		

Similarly, the third term represents the error induced by $\tilde{\mathbf{X}}^+$, and vanishes if $\tilde{\mathbf{x}}_i^+ = 0$. Also, $\lim_{N(k)\to\infty} \frac{1}{N(k)} \sum_{i=1}^{N(k)} \tilde{\mathbf{x}}_i^+ = 0$ if there is the finite number of indices such that $\tilde{\mathbf{x}}_i^+ \neq 0$.

The last term indicates the error induced by the current input. This term is zero as long as $\hat{\mathbf{x}}_{k-1} = \mathbf{x}_{k-1}$.

4.6 Simulation

We, in this section, present the simulations on the IEEE 68-bus test system (Figure 4.2) to demonstrate the performance of the ArGPR algorithm. The estimation results of the ArGPR algorithm are compared with those of the GP-EKF algorithm in [43, 124].

System model. We consider a power network represented by an undirected graph $(\mathcal{V}, \mathcal{E})$ where $\mathcal{V} \triangleq \{1, \cdots, 68\}$ and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ are the set of buses and the set of transmission lines, respectively. Let us denote $\mathcal{S}_i \triangleq \{l \in \mathcal{V} \setminus \{i\} | (i,l) \in \mathcal{E}\}$ the set of neighboring buses of $i \in \mathcal{V}$. Each bus is either a generator bus $i \in \mathcal{G}$ or a load bus $i \in \mathcal{L}$ and $\mathcal{V} = \mathcal{G} \cup \mathcal{L}$. The dynamic system of a generator bus $i \in \mathcal{G}$ with attacks is described as the following nonlinear system [80]:

$$\begin{aligned}
\Delta\dot{\theta}_i(t) &= \Delta f_i(t) + w_{1,i}(t) \\
\Delta\dot{f}_i(t) &= -\frac{1}{m_i}\Big(D_i \Delta f_i(t) + \sum_{j\in\mathcal{S}_i} \Delta P_{ij}(t) - \Delta P_{M_i}(t) + d_{a,i}(t) + \Delta P_{L_i}(t)\Big) + w_{2,i}(t) \\
y_{i,k} &= [\Delta\theta_{i,k}, \Delta f_{i,k}, \Delta P_{elec_i,k}]^T + [0, d_{s,i,k}^T]^T + v_{i,k}
\end{aligned} \tag{4.30}$$

where $\Delta P_{ij}(t) = t_{ij} \sin(\Delta\theta_i(t) - \Delta\theta_j(t))$ and $\Delta P_{elec_i,k} = \Delta P_{L_i}(t) + D_i \Delta f_i(t)$. Ta-

ble 4.1 summarizes the system variables and parameters and Δ denotes the distance from nominal value. Vectors $d_{a,i}(t) \in \mathbb{R}$ and $d_{s,i,k} \in \mathbb{R}^2$ denote actuator attack and sensor attack, respectively. The dynamic system for $i \in \mathcal{L}$ is

$$\begin{aligned}
\Delta\dot{\theta}_i(t) &= \Delta f_i(t) + w_{1,i}(t) \\
\Delta\dot{f}_i(t) &= -\frac{1}{m_i}\Big(D_i\Delta f_i(t) + \sum_{j\in\mathcal{S}_i}\Delta P_{ij}(t) + \Delta P_{C_i}(t) + d_{a,i}(t) + \Delta P_{L_i}(t)\Big) + w_{2,i}(t) \\
y_{i,k} &= [\Delta\theta_{i,k}, \Delta f_{i,k}, \Delta P_{elec_i,k}]^T + [0, d_{s,i,k}^T]^T + v_{i,k}. \qquad (4.31)
\end{aligned}$$

We will use functions F_i and H_i to express system (4.30) compactly; i.e., (4.30) becomes

$$\begin{aligned}
\dot{x}_i(t) &= F_i(x_i(t), [\Delta P_{M_i}(t), \{\Delta\theta_j(t)\}_{j\in\mathcal{S}_i}]) + w_i(t) \\
y_{i,k} &= H_i(x_{i,k}) + v_{i,k}
\end{aligned}$$

where $x_i(t) \triangleq [\Delta\theta_i(t), \Delta f_i(t)]^T$.

We assume that power demand $\Delta P_{L_i}(t)$ is known because it can be predicted by load forecasting methods [83,84]. Mechanical power $\Delta P_{M_i}(t)$ and controllable load $\Delta P_{C_i}(t)$ are considered known inputs of each bus, and we implement backstepping based stabilizing distributed controllers [106] for frequency control.

Simulation settings. Noises $w_i(t)$ and $v_{i,k}$ are zero-mean Gaussian with covariance $Q_i(t) = 0.01^2 I$, and $R_{i,k} = 0.01^2 I$. Sampling period is $\epsilon = 0.1s$. The system parameters are adopted from page 598 in [105], where $D_i = 1$, $t_{ij} = 1.5$, and $m_i = 10$ for $\forall i \in \mathcal{V}$.

We consider a scenario where the systems (4.30) and (4.31) for $\forall i$ are subject to both actuator attacks $d_{a,i}(t) = 30\sin(\frac{i\cdot t}{\pi}) + \frac{i}{100}$ for $t > 1$ and sensor attacks $d_{s,i,k} = [2 + 0.3\cos(0.1i \cdot t), 0]^T$ for $t > 7$.

Distributed implementation. The power system includes 204 sensors. The centralized implementation of the ArGPR algorithm requires 2^{204} modes, which is not practical. To address this problem, we implement the ArGPR algorithm in a distributed way.

Each bus is associated with a local defender. At time k, each local defender i measures $y_{i,k}$ and receives $\Delta\hat{\theta}_{j,k-1}$ from $j \in \mathcal{S}_i$. For the ArGPR algorithm, mechanical power $\Delta P_{M_i}(t)$ (controllable load $\Delta P_{C_i}(t)$, resp.) as well as the es-

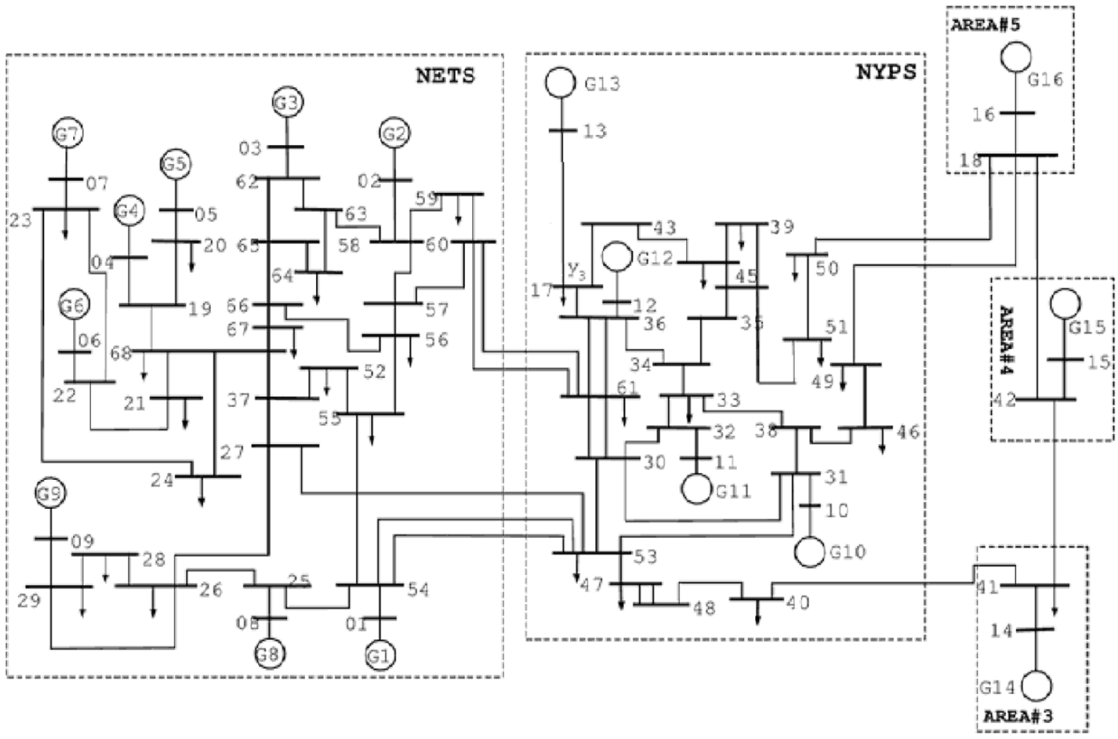

Figure 4.2: IEEE 68-bus test system (Figure from [1]).

timate of neighboring states $\hat{\theta}_{j,k}$ are treated as inputs for $i \in \mathcal{G}$ ($i \in \mathcal{L}$); i.e., $u_{i,k} = [\Delta P_{M_i,k}, \{\Delta\hat{\theta}_{j,k}\}_{j\in\mathcal{S}_i}]^T$ ($u_{i,k} = [\Delta P_{C_i,k}, \{\Delta\hat{\theta}_{j,k}\}_{j\in\mathcal{S}_i}]^T$). Each local defender is assumed to know the true hypothetical mode. We use only 1 randomly chosen training pair for the initial estimation. The GP-EKF algorithm in [43, 124] is implemented with the same settings for comparison.

Simulation results. Figure 4.3 summarizes the simulation results. In the two subfigures on the left hand side of Figure 4.3, the bold line represents the ArGPR algorithm, the dashed line represents the GP-EKF algorithm, and the solid thin line represents the ground truth. The first subfigure shows that state estimation errors of the ArGPR algorithm diminish, while those of the GP-EKF algorithm remain large. The second subfigure shows that frequency estimates of bus 16 converge to the ground truth. On the other hand, the GP-EKF algorithm fails to estimate states attack-resiliently. The right hand side of Figure 4.3 presents attack vector estimates $d_{1,i,k} = d_{s,i,k}$ and $d_{2,i,k} = [0, -d_{a,i}(t_k)/m_i]^T = [0, -3\sin(\frac{i\cdot t}{\pi}) - \frac{i}{1000}]^T$, and function approximation errors $\tilde{f}_{16,k}$ of bus 16. Attack vector estimates follow sinusoidal signals. Function approximation errors tend to decrease over time.

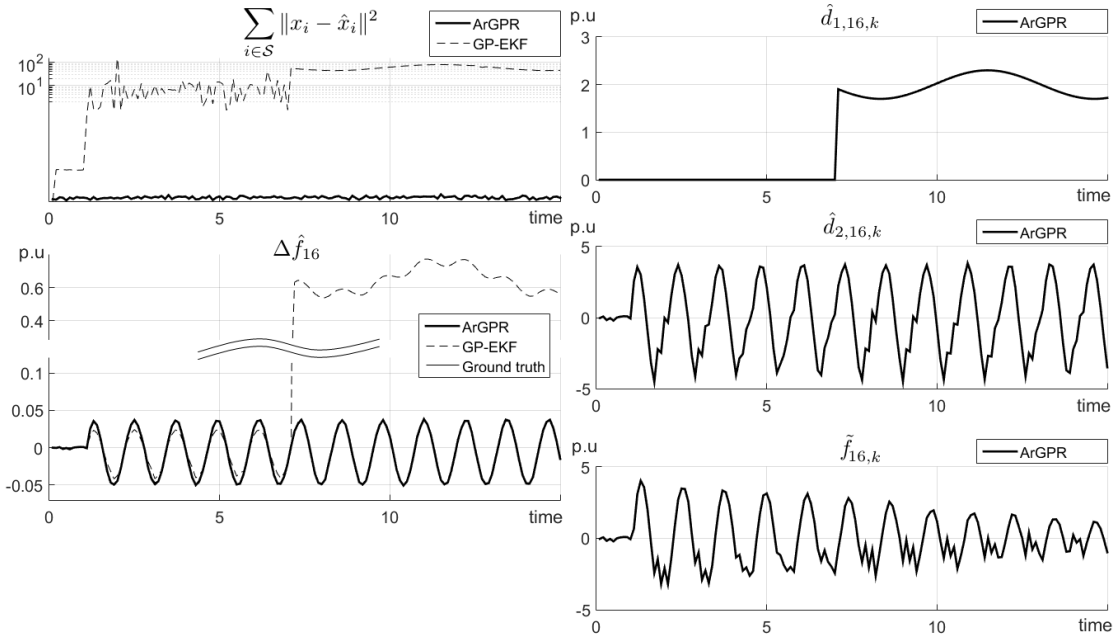

Figure 4.3: State estimation errors $\sum_{i\in\mathcal{V}} \|x_i - \hat{x}_i\|^2$ in log-scale, where $x_i = [\Delta\theta_i, \Delta f_i]^T$; frequency estimates of bus 16; attack vector estimates $\hat{d}_{1,16,k}$; the first element of attack vector estimates $\hat{d}_{2,16,k}$; and function approximation errors $\tilde{f}_{16,k}$ of bus 16.

4.7 Discussion

This section discusses comparisons with the most recent literature, and future works.

Comparison. The current chapter deals with partially known dynamic systems, where the system function is unknown but output function is known. This is the sharp difference from Chapter 2, which assumes perfect knowledge on systems. With the difference that follows, the current chapter incorporates a regression technique to solve the new problem, and analyze function approximation errors.

If attack vectors $d_{a,k}$ and $d_{s,k}$ are absent, the ArGPR algorithm reduces to the GP-EKF algorithm [43, 124] with known output equation. When system model is known, the ArGPR algorithm becomes a modification of the nonlinear unknown input and state estimation algorithm (Algorithm 1 in Chapter 2), where outputs are decomposed into three parts instead of two.

Mode convergence. Theorems 4.5.1 and 4.5.2 prove stability of estimation

errors and average case learning errors of function approximations under the assumption $|\mathcal{J}| = 1$. To relax the assumption $|\mathcal{J}| = 1$, convergence analysis of mode estimator (lines 5-7 in Algorithm 6) must be preceded because failing mode estimation degrades the performance of estimations and approximations.

Relaxing Assumption 4.5.1. The assumption implies that all components of actuator attack $d_{2,k-1}$ is measured by clean output $y^j_{2,k}$. Assumption 4.5.1 is a sufficient condition of that used in the unknown input and output estimation for fully-known linear systems (Theorem 5 in [89]). Assumption 4.5.1 can be relaxed to a weaker one by replacing $d_{2,k-1}$ with $G_{k-2}d_{2,k-1}$ in (4.10).

4.8 Conclusion

In this chapter, we study attack-resilient estimation of unknown nonlinear cyber-physical systems against both sensor attacks and actuator attacks. To solve the problem, we propose a new estimation algorithm by incorporating our recently developed unknown input and state estimation technique into the Gaussian process regression algorithm. We empirically demonstrate that the proposed algorithm estimates internal state attack-resiliently, outperforming the GP-EKF algorithm. Unlike existing attack detectors, the proposed algorithm does not require system models.

4.9 Appendix: GPR from linear regression

The objective of this appendix is to provide intuitions of GPR by deriving GPR from linear regression problem. This appendix is written based on [2].

Consider linear regression problem with Gaussian noise

$$\mathbf{z} = f(\mathbf{x}) + w, \quad f(\mathbf{x}) = \mathbf{x}^T\mathbf{a} \tag{4.32}$$

where $\mathbf{z} \in \mathbb{R}$, and $\mathbf{x} \in \mathbb{R}^n$. Vector $\mathbf{a} \in \mathbb{R}^n$ is a weight of the linear function f. Noise w follows independently identically distributed Gaussian distribution $w \sim \mathcal{N}(0, \sigma^2)$.

The objective of the linear regression is to find the posterior distribution of the weight vector $p(\mathbf{a}|D)$ by Baysian inference, given a set of input-output pair

observations $D = \langle \mathbf{X}, \mathbf{Z} \rangle$, called *training data set*, where

$$\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N],\ \mathbf{Z} = [\mathbf{z}_1, \cdots, \mathbf{z}_N].$$

The posterior distribution $p(\mathbf{a}|D)$ can be found by Bayes's rule

$$p(\mathbf{a}|D) = \frac{p(\mathbf{Z}|\mathbf{X}, \mathbf{a})p(\mathbf{a})}{p(\mathbf{Z}|\mathbf{X})}$$

where

$$p(\mathbf{Z}|\mathbf{X}, \mathbf{a}) = \mathcal{N}(\mathbf{X}^T\mathbf{a}, \sigma^2 I),\ p(\mathbf{Z}|\mathbf{X}) = \int p(\mathbf{Z}|\mathbf{X}, \mathbf{a})p(\mathbf{a})d\mathbf{a}$$

if prior distribution $p(\mathbf{a})$ is given. In GPR, we assume

$$\mathbf{a} \sim \mathcal{N}(0, \Sigma) \tag{4.33}$$

where Σ is the covariance matrix. This assumption suffices that $f(\mathbf{x})$ is GP. Under the assumption, we have

$$p(\mathbf{a}|\mathbf{D}) \sim \mathcal{N}(\frac{1}{\sigma^2}(\frac{\mathbf{X}\mathbf{X}^T}{\sigma^2} + \Sigma^{-1})^{-1}\mathbf{X}\mathbf{Z}, (\frac{\mathbf{X}\mathbf{X}^T}{\sigma^2} + \Sigma^{-1})^{-1}).$$

For a given testing input $\mathbf{x}_*$, output $f_* = f(\mathbf{x})$ follows

$$\begin{aligned} p(f_*|\mathbf{x}_*, D) &= \int p(f_*|x_*, \mathbf{a})p(\mathbf{a}|D)d\mathbf{a} \\ &= \mathcal{N}(\frac{1}{\sigma^2}\mathbf{x}_*^T(\frac{\mathbf{X}\mathbf{X}^T}{\sigma^2} + \Sigma^{-1})^{-1}\mathbf{X}\mathbf{Z}, \mathbf{x}_*^T(\frac{\mathbf{X}\mathbf{X}^T}{\sigma^2} + \Sigma^{-1})^{-1}\mathbf{x}_*). \end{aligned}$$

Feature space. The approach used for (4.34) is limited to a class of linear functions. In (4.34), we had applied weights to the input vector $\mathbf{x}$ directly. A simple extension is to apply weights to a set of basis functions $\phi(\mathbf{x})$ (feature vector) instead of directly to $\mathbf{x}$; i.e., function f is expressed as

$$f(\mathbf{x}) = \phi(\mathbf{x})^T\mathbf{a}$$

where $\phi : \mathbb{R}^n \to \mathbb{R}^m$. For example, $\phi(\mathbf{x}) = [1, \mathbf{x}, \mathbf{x}^2, log\mathbf{x}]^T$. A natural question is how to choose basis function ϕ. We will revisit this question later.

Under the assumption that $\mathbf{a}$ follows zero mean Gaussian prior (4.33), we can follow a similar procedure to obtain posterior distribution of $\mathbf{a}$:

$$p(\mathbf{a}|\mathbf{D}) \sim \mathcal{N}(\frac{1}{\sigma^2}(\frac{\Phi(\mathbf{X})\Phi(\mathbf{X})^T}{\sigma^2} + \Sigma^{-1})^{-1}\Phi(\mathbf{X})\mathbf{Z}, (\frac{\Phi(\mathbf{X})\Phi(\mathbf{X})^T}{\sigma^2} + \Sigma^{-1})^{-1})$$

where $\Phi(\mathbf{X})$ is a collection of $\phi(\mathbf{x}_i)$ for $i = 1, \cdots, N$. For a given testing input $\mathbf{x}_*$, output $f_* = f(\mathbf{x})$ follows

$$\begin{aligned}
&p(f_*|\mathbf{x}_*, D) \\
&= \mathcal{N}(\frac{1}{\sigma^2}\phi(\mathbf{x}_*)^T(\frac{\Phi(\mathbf{X})\Phi(\mathbf{X})^T}{\sigma^2} + \Sigma^{-1})^{-1}\Phi(\mathbf{X})\mathbf{Z}, \phi(\mathbf{x}_*)^T(\frac{\Phi(\mathbf{X})\Phi(\mathbf{X})^T}{\sigma^2} + \Sigma^{-1})^{-1}\mathbf{x}_*).
\end{aligned}$$

With new variables $g(x, x') = \phi(x)^T\Sigma\sigma(x')$ and $G(\mathbf{X}, \mathbf{X}') = \Phi(\mathbf{X})^T\Sigma\Phi(\mathbf{X}')$, the above posterior distribution becomes

$$\begin{aligned}
&p(f_*|\mathbf{x}_*, D) = \mathcal{N}(\phi(\mathbf{x}_*)^T\Sigma\Phi(\mathbf{X})(G(\mathbf{X}, \mathbf{X}) + \sigma^2 I))^{-1}\mathbf{Z}, \\
&- \phi(\mathbf{x}_*)^T\Sigma\Phi(\mathbf{X})(G(\mathbf{X}, \mathbf{X}) + \sigma^2 I))^{-1}\Phi(\mathbf{X})^T\Sigma\phi(\mathbf{x}_*)) \\
&= \mathcal{N}(g_*(G(\mathbf{X}, \mathbf{X}) + \sigma^2 I))^{-1}\mathbf{Z}, g(x_*) - g_*(G(\mathbf{X}, \mathbf{X}) + \sigma^2 I))^{-1}g_*^T) \qquad (4.34)
\end{aligned}$$

where $g_* = G(x_*, \mathbf{X})\phi(\mathbf{x}_*)^T\Sigma\Phi(\mathbf{X})$. Please note that (4.34) is identical to GPR regression obtained in (4.3) and (4.4). Function $g(x, x') = \phi(x)^T\Sigma\phi(x')$ is called a kernel. Since Σ is positive definite, we can express the kernel as a dot product $g(x, x') = \psi(x)^T\psi(x')$ where $\psi(x) = \Sigma^{1/2}\phi(x)$.

The posterior distribution is found by inner products in input space, and we can replace the basis function ϕ by changing kernel g. This is called the kernel trick. In the kernel trick, choosing basis function ϕ is equivalent to choosing corresponding kernel g. Thus, one can simply choose a positive definite kernel for GPR without explicitly investigate a set of basis functions. Moreover, a kernel potentially exhibits an infinite dimensional basis functions.

Chapter 5
Incentive design

5.1 Introduction

Advanced information and communication technologies have been stimulating rapid emergence of multi-agent networks where a large number of spatially distributed agents interact with each other to accomplish complex missions. Substantial effort has been spent on analysis, design and control of multi-agent networks [5–9]. In many practical scenarios, agents are non-cooperative and seek for heterogeneous (or even conflicting) subobjectives. This leads to competitions over limited resources and the degradation of network-wide performance. To mitigate the issue, a common practice is incentive or mechanism design which modifies agents' preferences via side payments/pricing so that individual interests are aligned with social welfare.

As one kind of incentive design, fixed prize lotteries have been applied to several field experiments and proven to be effective to stimulate agents' or players' investments. INSINC project in Singapore [125] is an ongoing real-world implementation of a lottery scheme for commuters, who use public transportation, to travel off peak hours. The lottery scheme successfully reduces around 7.5% of peak time demand. A similar project named INSTANT [126] is conducted in India and results in more than 20% of commuter shifts. Research [127] uses the boarding passes of local public transportation as lottery tickets, showing that the lottery increases the provision of public goods and reduces free riders. In [128–130], experiments are conducted to show that lottery based incentives can effectively increase survey response rates. Moreover, lottery based incentives have been used

in demand response in the smart grid [131, 132], mobile crowd sensing for traffic congestion and air pollution [133] and Internet congestion [134].

Substantial effort has been exerted to develop fundamental theory of lotteries. Seminal paper [60] studies that fixed prize lotteries alleviate the free-rider problem, and nudge higher levels of public good provisions as well as aggregate payoff than voluntary contributions. A larger reward results in a greater public good and aggregate payoff. The results have been extended by many researchers. In [135], a multi-prize lottery is studied with considering risk preferences; i.e., risk neutral versus risk averse. A sequential lottery is investigated in [136] in which it can sale more tickets than one-level lottery. Paper [137] conducts public good analysis on player size, and extends the results to a rival public good case; i.e., each player gets benefits from a portion of public goods.

Chapter organization. In Section 5.2, we introduce a classic bi-level lottery scheme with its limitation. To alleviate the fundamental limitation of efficiency losses, we introduce a new perturbed bi-level lottery model and formulate the optimal bi-level lottery design problem in Section 5.3. In Section 5.4, we formally analyze properties of low-level Nash equilibrium. Based on the properties, we relax the optimal bi-level lottery design problem as a convex optimization problem in Section 5.5. Lastly, we conduct a case study on demand response in Section 5.6.

5.2 Preliminaries

We introduce a classic bi-level lottery scheme proposed in [60] and outline its procedure in Section 5.2.1 to 5.2.3. Section 5.2.4 discusses its limitation and motivates our problem.

5.2.1 Payoff model

Consider a social planner who holds a lottery and a set of players $\mathcal{V} = \{1, 2, \cdots, N\}$. In particular, the social planner chooses a reward R from an action set $\mathcal{R} = (0, \infty)$. Each player i invests s_i to the lottery from an action set $\mathcal{S}_i = [0, w_i]$ and is associated with a payoff function $u_i : \mathcal{S} \to \mathbb{R}$ where w_i denotes the amount of investable wealth of player i, and $\mathcal{S} \triangleq \mathcal{S}_1 \times \cdots \times \mathcal{S}_N$ denotes the joint action set. The action profile $s = \{s_i\}_{i \in \mathcal{V}} \in \mathcal{S}$ can be expressed as $\{s_i, s_{-i}\}$ where s_{-i} denotes the action

profile other than player i; i.e., $s_{-i} = \{s_j\}_{j\in\mathcal{V}\setminus\{i\}}$. Given reward R, payoff function u_i is described by:

$$u_i(s,R) \triangleq \begin{cases} \frac{s_i}{\bar{s}}R + h_i(\bar{s}-R) - s_i, & \text{for } \bar{s} \geq R \\ 0, & \text{otherwise} \end{cases} \tag{5.1}$$

where $\bar{s} \triangleq \sum_{i\in\mathcal{V}} s_i$. If $\bar{s} = 0$, then $u_i(s,R) = 0$. The first term $\frac{s_i}{\bar{s}}R$ represents the portion of profit from the lottery, which is proportional to investment over the total investment. The last term $-s_i$ denotes the cost of player i. Marginal benefit function $h_i : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ satisfies the following:

Assumption 5.2.1 *The function h_i is twicely differentiable, strictly increasing, strictly concave, $h_i(0) = 0$, $\sum_{i\in\mathcal{V}} \frac{\partial h_i(0)}{\partial G} > 1$, and $\lim_{G\to\infty} \frac{\partial h_i(G)}{\partial G} = 0$.*

Payoff function (5.1) indicates that the lottery holds only when total investment $\bar{s}$ exceeds or equals to reward R; otherwise, the social planner cancels the lottery and returns the investments to the players.

5.2.2 Low-level decision making - Nash equilibrium

Given R and s_{-i}, player i chooses s_i to maximize its own payoff as follows:

$$\max_{s_i\in\mathcal{S}_i} u_i(s,R).$$

The collection of local optimization problems induces a non-cooperative game among the players and the game is parameterized by R. Nash equilibrium [138] defines the solution of the game.

Definition 5.2.1 *Given R, the action profile $s^*(R)$ is a (pure) Nash equilibrium if $u_i(s_i', s^*_{-i}(R), R) \leq u_i(s^*(R), R)$ for $\forall s_i' \in \mathcal{S}_i, \forall i \in \mathcal{V}$.*

Note that Nash equilibrium $s^*(R)$ highlights its dependency on reward R.

5.2.3 High-level decision making - Social optimum

The lottery is a bi-level decision making (or a hierarchical optimization) problem where the social planner at the high level selects reward first and, sequentially, the players at the low level jointly determine a Nash equilibrium given the reward.

The social planner aims to choose reward R to maximize the aggregate payoff of the players at the induced Nash equilibrium:

$$\max_{R\in\mathcal{R}} \sum_{i\in\mathcal{V}} u_i(s^*(R), R) = \max_{R\in\mathcal{R}} \sum_{i\in\mathcal{V}} h_i(G(R)) - G(R) \tag{5.2}$$

where $G(R) \triangleq \bar{s}^*(R) - R$ is referred to as the *public good* and represents a marginal benefit of the social planner. The hierarchical nature of the problem requires the social planner to predict the low-level Nash equilibrium when making decisions at the high level.

5.2.4 Limitation

Under Assumption 5.2.1, there exists a unique socially optimal public good (Proposition 2.1 in [60])

$$G^* = \text{argmax}_{G\in[0,\infty)} \sum_{i\in\mathcal{V}} h_i(G) - G,$$

where $G^* > 0$ is the solution of

$$\sum_{i\in\mathcal{V}} \frac{\partial h_i(G^*)}{\partial G} = 1 \tag{5.3}$$

due to strict concaveness of h_i. The socially optimal public good maximizes the aggregate payoff, and we define the aggregate payoff

$$\sum_{i\in\mathcal{V}} h_i(G^*) - G^*$$

as the *socially optimal payoff*. However, the socially optimal public good (as well as socially optimal payoff) is achieved only when $R \to \infty$ (Theorem 2 in [60]). An infinite reward is apparently impractical. The chapter aims to address the limitation.

5.3 Problem formulation

This section introduces a practical scheme to achieve the socially optimal payoff. In particular, a perturbed lottery model is introduced in Section 5.3.1 and lower-level decision-making is presented in Section 5.3.2. A new problem for the social planner is introduced in Section 5.3.3. We highlight the differences from those introduced in Section 5.2.

5.3.1 Perturbed payoff model

Consider the following *perturbed payoff model* for player i:

$$U_i(s,R,c) \triangleq \begin{cases} \frac{s_i-c_i}{\bar{s}-\bar{c}}R + h_i(\bar{s}-R) - \beta_i s_i, & \text{for } \bar{s} \geq R \\ 0, & \text{otherwise} \end{cases} \tag{5.4}$$

where β_i and c_i are *heterogeneity parameter* and *perturbation parameter*, respectively, with $c = \{c_i\}_{i\in\mathcal{V}}$ and $\bar{c} = \sum_{i\in\mathcal{V}} c_i$. Heterogeneity parameter $0 < \beta_i \leq 1$ is determined by player i before playing the game. It represents valuation on investments, or willingness of voluntary investments.

Assumption 5.3.1 *The parameters β_i for $\forall i \in \mathcal{V}$ satisfy $N-1 < \bar{\beta} \leq N$, and $0 < \beta_i \leq 1$.*

In (5.4), (R,c) is chosen by the social planner from a set $\mathcal{R}\times\mathcal{C}$ where $\mathcal{C} \triangleq \mathcal{C}_1 \times \cdots \times \mathcal{C}_N$, and $\mathcal{C}_i = [0,\infty)$. Perturbation parameter c introduces an offset to the odd of winning but the aggregate portion remains one; i.e., $\sum_{i\in\mathcal{V}} \frac{s_i-c_i}{\bar{s}-\bar{c}} = 1$. We show later in (P2) of Theorem 5.4.1 that $\bar{s} \neq \bar{c}$.

5.3.2 Low-level decision making - Nash equilibrium

Given R, c, and s_{-i}, player i chooses s_i to maximize its own payoff as follows:

$$\max_{s_i\in\mathcal{S}_i} U_i(s,R,c),$$

where $\mathcal{S}_i = [0,\infty)$. Nash equilibrium $s^*(R,c)$ is dependent on R and c. If the reward term $\frac{s_i^*(R,c)-c_i}{\bar{s}^*(R,c)-\bar{c}}R$ is negative, player i is assumed to pay a fine $-\frac{s_i^*(R,c)-c_i}{\bar{s}^*(R,c)-\bar{c}}R$ to the social planner.

5.3.3 High-level decision making - Social optimum

As problem (5.2), a natural problem for the social planner is to maximize the aggregate of perturbed payoffs as follows:

$$\max_{(R,c)\in\mathcal{R}\times\mathcal{C}} \sum_{i\in\mathcal{V}} (h_i(G(R,c)) + (1-\beta_i)s_i^*(R,c)) - G(R,c). \tag{5.5}$$

For the case with $\beta_i = 1$ for $\forall i$, there could be multiple optimal solutions for problem (5.5). So the social planner may want to choose the one induced by minimal reward and perturbation. For the case with $\beta_i \neq 1$ for some i, problem (5.5) is not well-defined. In particular, we will show in (P3) of Theorem 5.4.1 that $s_i^*(R,c)$ approaches infinity as R goes to infinity. For this case, the social planner may want to choose a pair of small R and c such that the induced aggregate payoff is not smaller than the socially optimal payoff of problem (5.2). With the above two cases, problem (5.5) is reformulated as the following bi-level optimization problem:

$$\begin{aligned} &\min_{(R,c)\in\mathcal{R}\times\mathcal{C}} R + \alpha\bar{c} \\ &\text{s.t.} \quad g(s^*(R,c),R,c) \leq 0,\ \bar{c} \leq \min\{G^U, R\} \\ &\qquad \sum_{i\in\mathcal{V}} h_i(G(R,c)) - G(R,c) + \sum_{i\in\mathcal{V}} (1-\beta_i)s_i^*(R,c) \geq \sum_{i\in\mathcal{V}} h_i(G^*) - G^* \end{aligned} \tag{5.6}$$

where constant $\alpha \geq 0$ represents a weight on $\bar{c}$, and the dependency of public good $G(R,c) \triangleq \bar{s}^*(R,c) - R$ on R and c is emphasized. The constraint $\bar{c} \leq \min\{G^U, R\}$ is imposed due to a technical reason, and the value G^U is the solution of $\sum_{i\in\mathcal{V}} \frac{\partial h_i(G^U)}{\partial G} = \bar{\beta} + 1 - N$. We include a new inequality constraint $g(s^*, R, c) \leq 0$ where $g : \mathcal{S}\times\mathcal{R}\times\mathcal{C} \to \mathbb{R}^m$ is a vector of convex functions $g_\ell(s^*, R, c)$ for $\ell = 1, 2, \cdots, m$. The new constraint might express physical constraints as shown in Section 5.6, or planner's interests.

Assumption 5.3.2 *Function $g_\ell(s^*, R, c)$ is convex with respect to its arguments s^*, R, and c for $\ell = 1, 2, \cdots, m$.*

To clarify the relation between problem (5.2) and problem (5.6), let us distinguish two cases.

Case 1. $\beta_i = 1$ for $\forall i$.

When the first two constraints are absent, problem (5.6) returns the pair of minimal

R and c which can induce G^* as well as the socially optimal payoff of problem (5.2).

Case 2. $\beta_i \neq 1$ for some i.

Since $(1-\beta_i)s_i^*(R,c) \geq 0$, the optimal solution of problem (5.6) may not induce G^*. Yet, it is still able to induce an aggregate payoff which is not smaller than the socially optimal payoff of problem (5.2).

Notice that the feasible set of problem (5.6) is always non-empty when the new constraint $g(s^*(R,c),R,c) \leq 0$ is not considered in the unperturbed lottery. The proof of the following lemma is presented in Appendix 5.8.

Lemma 5.3.1 *Under Assumptions 5.2.1, and 5.3.1, the feasible set of problem* (5.6) *is non-empty if constraint* $g(s^*(R,c),R,c) \leq 0$ *is absent.*

5.4 Analysis of low-level Nash equilibrium

In this section, we study the properties of Nash equilibrium given a pair of R and c which satisfies Assumption 5.4.1.

Assumption 5.4.1 *Social planner's action pair* $(R,c) \in \mathcal{R} \times \mathcal{C}$ *satisfies* $\bar{c} \leq \min\{G^U, R\}$.

Theorem 5.4.1 summarizes the derived properties, and these properties are essential to solve bi-level optimization problem (5.6). Further, the properties reduce to those of unperturbed lottery in Section 5.2 when $c_i = 0$ and $\beta_i = 1$. In particular, (P1) shows the existence and uniqueness of Nash equilibrium. (P2) indicates that public good $G(R,c)$ is lower bounded by a function of c and upper bounded by a function of β, and it is increasing in (R,c) when all the players are active. (P3) shows that all the players are active if reward R is greater than a certain threshold, and there exists a lower bound of $s_i^*(R,c)$, which is a strictly increasing function in R. Moreover, in some cases, $s_i^*(R,c)$ is strictly increasing in (R,c). (P4) quantifies the price of anarchy [139] which is the ratio between the socially optimal payoff and the aggregate payoff induced by the corresponding Nash equilibrium. The lower and upper bounds of the price of anarchy reveal possible efficiency losses due to selfishness of players, and they can be quantified without explicitly calculating Nash equilibrium.

The following notations are used in Theorem 5.4.1. The value $R_L(c)$ is the unique solution of $\frac{R_L(c)}{R_L(c)+G^U-\bar{c}} = \max_{i\in\mathcal{N}}\{\beta_i - \frac{\partial h_i(G^U)}{\partial G}\}$. Define player i who in-

vests non-zero wealth $s_i^*(R,c) > 0$ as an active player and define $\mathcal{V}_a(R,c) \triangleq \{i \in \mathcal{V} | s_i^*(R,c) > 0\}$ as the set of all the active players. Lastly,

$$\underline{G}(R,c) \triangleq H^{-1}(\frac{(|\bar{\mathcal{V}}_a(R,c)|-1)(\bar{G}(R,c)-\bar{c})}{R+G^U-\bar{c}}+1)$$
$$\bar{G}(R,c) \triangleq H^{-1}(\frac{(N-1)(G^*-\bar{c})}{R}+1)$$

where $H(G) \triangleq \sum_{i\in\mathcal{V}} \frac{\partial h_i(G)}{dG}$ and $\bar{\mathcal{V}}_a(R,c)$ is the number of players who satisfy $\frac{R}{R+G^U-\bar{c}} + \frac{\partial h_i(G^U)}{\partial G} - 1 > 0$. Note that $H : \mathbb{R}_{\geq 0} \to Y$ is invertible on codomain $Y \triangleq (0, H(0)]$ because H is a strictly decreasing and continuous.

Theorem 5.4.1 *Under Assumptions 5.2.1, 5.3.1, and 5.4.1, the following properties hold at Nash equilibrium.*

(P1) *Given any $c_i \geq 0$, and $0 < \beta_i \leq 1$ for $\forall i \in \mathcal{V}$, there is a unique Nash equilibrium $s^*(R,c)$;*

(P2) *It holds that $\bar{c} \leq G(R,c) \leq G^U$. If $|\mathcal{V}_a(R,c)| = N$, then $\frac{dG(R,c)}{dR} \geq 0$, and $\frac{dG(R,c)}{dc_i} > 0$ where equality holds if and only if $\bar{c} = G^U$;*

(P3) *If $R > R_L(c)$, then $s_i^*(R,c) \geq c_i + R\Big(\frac{R}{R+G^U-\bar{c}} + \frac{\partial h_i(G^U)}{\partial G} - \beta_i\Big) > 0$ where the lower bound is strictly increasing in R without bound. If $|\mathcal{V}_a(R,c)| = N$, there is some $i \in \mathcal{V}$ such that $\frac{ds_i^*(R,c)}{dR} > 0$. Moreover, if $|\mathcal{V}_a(R,c)| = N$, $h_i = h_j$, and $\beta_i = \beta_j$ for $\forall i,j \in \mathcal{V}$, then $\frac{ds_i^*(R,c)}{dR} \geq \frac{1}{N}$, and $\frac{ds_i^*(R,c)}{dc_i} > 0$ for $\forall i$;*

(P4) *If $\beta_i = 1$ for $\forall i \in \mathcal{V}$ and $R > 0$, price of anarchy $\mathrm{PoA}(R,c) \triangleq \frac{\max_{s\in\mathcal{S}} \sum_{i\in\mathcal{V}} U_i(s)}{\sum_{i\in\mathcal{V}} U_i(s^*(R,c),R,c)}$ is characterized by*

$$\frac{\sum_{i\in\mathcal{V}} h_i(G^*) - G^*}{\sum_{i\in\mathcal{V}} h_i(\underline{G}(R,c)) - \underline{G}(R,c)} \leq \mathrm{PoA}(R,c) \leq \frac{\sum_{i\in\mathcal{V}} h_i(G^*) - G^*}{\sum_{i\in\mathcal{V}} h_i(\bar{G}(R,c)) - \bar{G}(R,c)}.$$

If $c = \vec{0}$, it holds that $\mathrm{PoA} > 1$ for any $R < \infty$ and $\lim_{R\to\infty} \mathrm{PoA}(R,\vec{0}) = 1$.

PROOF. In the proof, we will drop the dependency of G, $\underline{G}$, $\bar{G}$, s^*, U_i, $\mathcal{V}_a$, R_L and PoA on R and c.

We first introduce the first order condition which must be satisfied at a Nash equilibrium:

$$\frac{\partial U_i(s^*, R, c)}{\partial s_i} = R\frac{\bar{s}^* - \bar{c} - (s_i^* - c_i)}{(\bar{s}^* - \bar{c})^2} + \frac{\partial h_i(\bar{s}^* - R)}{\partial G} - \beta_i \leq 0 \tag{5.7}$$

for $\forall i \in \mathcal{V}$. If player i is active; i.e., $s_i^* > 0$, then equality holds. To prove the first order condition by contradiction, assume that $\frac{\partial U_i(s^*,R,c)}{\partial s_i} = \epsilon > 0$. Then, by the Taylor series expansion, there exists a sufficiently small constant $\delta > 0$ such that

$$U_i(s_i^* + \epsilon\delta, s_{-i}^*, R, c) > U_i(s_i^*, s_{-i}^*, R, c) + \epsilon\delta.$$

This leads to a contradiction to the definition of Nash equilibrium. The remaining part can be proven in a similar way.

(P1) Choose any β. Since h_i is strictly increasing and strictly concave, there is $\xi_L > 0$ such that $\frac{\partial h_i(\xi)}{\partial G} < 1$ for all $\xi \geq \xi_L$. Consider any R, c and s_{-i}. If s_i is sufficiently large, then $U_i(s) < 0$. So there is $B_i(R, c) > 0$ such that $s_i^*(R, c) < B_i(R, c)$. Hence, $s^*(R, c)$ is identical to the maximizer of the game: $\max_{s_i} U_i(s)$ s.t. $s_i \in [0, B_i(R, c)]$. In this problem, the payoff functions are concave and the decision variables lie in compact sets. Hence, $s^*(R, c)$ exists. The remaining part can be proven by similar arguments of Lemma 3 in [60].

(P2) The aggregate of the first order conditions (5.7) becomes

$$\sum_{i\in\mathcal{V}} \frac{\partial U_i(s^*)}{\partial s_i} = \frac{R(N-1)}{R + G - \bar{c}} + \sum_{i\in\mathcal{V}} \frac{\partial h_i(G)}{\partial G} - \bar{\beta} \leq 0. \tag{5.8}$$

Assume $G < \bar{c}$, then (5.8) yields

$$\sum_{i\in\mathcal{V}} \frac{\partial h_i(G)}{\partial G} < \bar{\beta} + 1 - N = \sum_{i\in\mathcal{V}} \frac{\partial h_i(G^U)}{\partial G}.$$

This implies $G^U < G < \bar{c}$ due to strict concaveness of h_i, which contradicts Assumption 5.4.1, and thus $\bar{c} \leq G$. Now consider the aggregate of the first order conditions (5.7) of active players:

$$\sum_{i\in\mathcal{V}_a} \frac{\partial U_i(s^*)}{\partial s_i} = \frac{R(|\mathcal{V}_a| - 1)}{R + G - \bar{c}} + \frac{\sum_{i\in\mathcal{V}\setminus\mathcal{V}_a}(s_i^* - c_i)}{(R + G - \bar{c})^2}R + \sum_{i\in\mathcal{V}_a} \frac{\partial h_i(G)}{\partial G} - \sum_{i\in\mathcal{V}_a} \beta_i = 0.$$

Note that $s_i^* = 0$ for $i \in \mathcal{V} \setminus \mathcal{V}_a$. By the fact that h_i is a strictly increasing function, it becomes

$$\begin{aligned}
\sum_{i\in\mathcal{V}} \frac{\partial h_i(G)}{\partial G} &\geq -\frac{R(|\mathcal{V}_a|-1)}{R+G-\bar{c}} + \sum_{i\in\mathcal{V}_a} \beta_i + \frac{\sum_{i\in\mathcal{V}\setminus\mathcal{V}_a} c_i}{(R+G-\bar{c})^2} R \\
&= \frac{(|\mathcal{V}_a|-1)(G-\bar{c})}{R+G-\bar{c}} + \sum_{i\in\mathcal{V}_a} \beta_i + 1 - |\mathcal{V}_a| + \frac{\sum_{i\in\mathcal{V}\setminus\mathcal{V}_a} c_i}{(R+G-\bar{c})^2} R \\
&\geq \frac{(|\mathcal{V}_a|-1)(G-\bar{c})}{R+G-\bar{c}} + \sum_{i\in\mathcal{V}} \frac{\partial h_i(G^U)}{\partial G} + \frac{\sum_{i\in\mathcal{V}\setminus\mathcal{V}_a} c_i}{(R+G-\bar{c})^2} R. \qquad (5.9)
\end{aligned}$$

Because $\frac{(|\mathcal{V}_a|-1)(G-\bar{c})}{R+G-\bar{c}} \geq 0$ and $c_i \geq 0$, (5.9) implies $G \leq G^U$ by strict concaveness of h_i. Thus, $\bar{c} \leq G \leq G^U$.

Now we consider the case with $|\mathcal{V}_a| = N$. Since all the players are active the aggregate first order condition (5.8) holds with equality where $\sum_{i\in\mathcal{V}} \frac{\partial U_i(s^*)}{\partial s_i}$ can be regarded as an implicit function of (s^*, R, c). We apply the implicit function theorem (Theorem 1.3.1 in [140]) to (5.8)

$$-\frac{\partial(\sum_{i\in\mathcal{V}} \frac{\partial U_i(s^*)}{\partial s_i})}{\partial G} \frac{dG}{dR} = \frac{\partial(\sum_{i\in\mathcal{V}} \frac{\partial U_i(s^*)}{\partial s_i})}{\partial R}$$

and obtain

$$\frac{dG}{dR} = -\frac{(G-\bar{c})(N-1)}{(R+G-\bar{c})^2 \sum_{i\in\mathcal{V}} \frac{\partial^2 h_i(G)}{\partial G^2} - R(N-1)} \geq 0. \qquad (5.10)$$

It holds that $\frac{dG}{dR} = 0$ if and only if $G = \bar{c}$.

We will show that $G = \bar{c}$ if and only if $\bar{c} = G^U$. If $\bar{c} = G^U$, then $G = \bar{c}$ because $\bar{c} \leq G \leq G^U$. We now prove that if $G = \bar{c}$ then $\bar{c} = G^U$. Assume $G = \bar{c}$, then aggregate first order condition (5.8) yields

$$\sum_{i\in\mathcal{V}} \frac{\partial U_i(s^*)}{\partial s_i} = N - 1 + \sum_{i\in\mathcal{V}} \frac{\partial h_i(\bar{c})}{\partial G} - \bar{\beta} = 0.$$

The unique solution is $\bar{c} = G^U$.

We proceed to prove $\frac{dG}{dc_i} > 0$. By applying the implicit function theorem

to (5.8), we have

$$-\frac{\partial(\sum_{i\in\mathcal{V}}\frac{\partial U_i(s^*)}{\partial s_i})}{\partial G}\frac{dG}{dc_i}=\frac{\partial(\sum_{i\in\mathcal{V}}\frac{\partial U_i(s^*)}{\partial s_i})}{\partial c_i}$$

and obtain

$$\frac{dG}{dc_i}=-\frac{R(N-1)}{(R+G-\bar{c})^2\sum_{i\in\mathcal{V}}\frac{\partial^2 h_i(G)}{\partial G^2}-R(N-1)}>0. \tag{5.11}$$

(P3) By $G\le G^U$ and concaveness of h_i, first order condition (5.7) yields

$$\begin{aligned}\frac{\partial U_i(s^*)}{\partial s_i}&=R\frac{\bar{s}^*-\bar{c}-(s_i^*-c_i)}{(\bar{s}^*-\bar{c})^2}+\frac{\partial h_i(\bar{s}^*-R)}{\partial G}-\beta_i\\&\ge\frac{R}{\bar{s}^*-\bar{c}}-R\frac{s_i^*-c_i}{(\bar{s}^*-\bar{c})^2}+\frac{\partial h_i(G^U)}{\partial G}-\beta_i.\end{aligned} \tag{5.12}$$

Assume $s_i^*<c_i$, then with $R>R_L$,

$$\frac{\partial U_i(s^*)}{\partial s_i}>\frac{R_L}{R_L+G^U-\bar{c}}+\frac{\partial h_i(G^U)}{\partial G}-\beta_i=0.$$

This contradicts the first order condition, and thus $s_i^*\ge c_i$. With $G\ge\bar{c}$, (5.12) becomes

$$\frac{\partial U_i(s^*)}{\partial s_i}\ge\frac{R}{R+G^U-\bar{c}}-\frac{s_i^*-c_i}{R}+\frac{\partial h_i(G^U)}{\partial G}-\beta_i.$$

If $s_i^*<c_i+R\Big(\frac{R}{R+G^U-\bar{c}}+\frac{\partial h_i(G^U)}{\partial G}-\beta_i\Big)$, then $\frac{\partial U_i(s^*)}{\partial s_i}>0$, a contradiction to the first order condition. Therefore $s_i^*\ge c_i+R\Big(\frac{R}{R+G^U-\bar{c}}+\frac{\partial h_i(G^U)}{\partial G}-\beta_i\Big)$ and the lower bound is strictly positive, because $R>R_L$.

We now proceed to prove that the bound $L_i(R,c)\triangleq c_i+R\Big(\frac{R}{R+G^U-\bar{c}}+\frac{\partial h_i(G^U)}{\partial G}-\beta_i\Big)$ is a strictly increasing function of R without bound. By taking derivative of the bound, we have

$$\frac{\partial L_i}{\partial R}=\frac{R}{R+G^U-\bar{c}}+\frac{\partial h_i(G^U)}{\partial G}-\beta_i+R\frac{G^U-\bar{c}}{(R+G^U-\bar{c})^2}$$

which is strictly greater than 0 since $G^U\ge\bar{c}$ and $R>R_L$. Moreover, function L_i keeps increasing without bound as R increases because $\lim_{R\to\infty}\frac{\partial L_i}{\partial R}=1+\frac{\partial h_i(G^U)}{\partial G}-$

$\beta_i > 0$.

Now we will consider the case with $|\mathcal{V}_a| = N$. We will show that there is at least one i such that $\frac{ds_i^*}{dR} > 0$ holds. Since all the players are active, the first order condition (5.7) holds with equality $\frac{\partial U_i(s^*)}{\partial s_i} = 0$ where $\frac{\partial U_i(s^*)}{\partial s_i}$ can be regarded as an implicit function of (s^*, R, c). By the implicit function theorem, relation

$$-\begin{bmatrix} \frac{\partial^2 U_1(s^*)}{\partial s_1^2} & \cdots & \frac{\partial^2 U_1(s^*)}{\partial s_1 \partial s_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 U_N(s^*)}{\partial s_N \partial s_1} & \cdots & \frac{\partial^2 U_N(s^*)}{\partial s_N^2} \end{bmatrix} \begin{bmatrix} \frac{ds_1^*}{dR} \\ \vdots \\ \frac{ds_N^*}{dR} \end{bmatrix} = \begin{bmatrix} \frac{\partial^2 U_1(s^*)}{\partial s_1 \partial R} \\ \vdots \\ \frac{\partial^2 U_N(s^*)}{\partial s_N \partial R} \end{bmatrix}$$

holds where

$$\begin{aligned} \frac{\partial^2 U_i(s^*)}{\partial s_i^2} &= -2R\frac{\bar{s}^* - \bar{c} - (s_i^* - c_i)}{(\bar{s}^* - \bar{c})^3} + \frac{\partial^2 h_i(\bar{s}^* - R)}{\partial G^2} < 0 \\ \frac{\partial^2 U_i(s^*)}{\partial s_i \partial s_j} &= -R\frac{\bar{s}^* - \bar{c} - 2(s_i^* - c_i)}{(\bar{s}^* - \bar{c})^3} + \frac{\partial^2 h_i(\bar{s}^* - R)}{\partial G^2} \\ \frac{\partial^2 U_i(s^*)}{\partial s_i \partial R} &= \frac{\bar{s}^* - \bar{c} - (s_i^* - c_i)}{(\bar{s}^* - \bar{c})^2} - \frac{\partial^2 h_i(\bar{s}^* - R)}{\partial G^2} > 0. \end{aligned}$$

If we choose $k = \text{argmin}_{i \in \mathcal{V}}\ s_i^*$, it holds that $\frac{\partial^2 U_k(s^*)}{\partial s_k \partial s_j} \leq 0$ because $s_k^* > c_k$. Therefore, the relation

$$-\sum_{j \in \mathcal{V}} \frac{\partial^2 U_k(s^*)}{\partial s_k \partial s_j} \frac{ds_j^*}{dR} = \frac{\partial^2 U_k(s^*)}{\partial s_k \partial R} > 0$$

implies that there is at least one j such that $\frac{ds_j^*}{dR} > 0$.

We now proceed to prove the remaining part. Assume that $h_i = h_j$, and $\beta_i = \beta_j$ for $\forall i, j \in \mathcal{V}$. Since all the players are active and thus first order condition (5.7) yields

$$s_i^* = c_i + R + G - \bar{c} + \frac{(R + G - \bar{c})^2}{R}(\frac{\partial h_i(G)}{\partial G} - \beta_i). \tag{5.13}$$

Using the chain rule,

$$\frac{ds_i^*}{dR} = \frac{\partial s_i^*}{\partial R} + \frac{\partial s_i^*}{\partial G}\frac{dG}{dR}$$

where $\frac{\partial s_i^*}{\partial R}$ and $\frac{\partial s_i^*}{\partial G}$ can be found from (5.13), and $\frac{dG}{dR}$ is in (5.10). One can show that $\frac{ds_i^*}{dR} = \frac{ds_j^*}{dR}$. Since $\frac{dG}{dR} \geq 0$, we have

$$\frac{dG}{dR} = -1 + \sum_{i \in \mathcal{V}} \frac{ds_i^*}{dR} = -1 + N \frac{ds_i^*}{dR} \geq 0.$$

Therefore, $\frac{ds_i^*}{dR} \geq \frac{1}{N}$.

The chain rule yields

$$\frac{ds_i^*}{dc_i} = \frac{\partial s_i^*}{\partial c_i} + \frac{\partial s_i^*}{\partial G}\frac{dG}{dc_i}, \quad \frac{ds_j^*}{dc_i} = \frac{\partial s_j^*}{\partial c_i} + \frac{\partial s_j^*}{\partial G}\frac{dG}{dc_i}.$$

where $\frac{\partial s_i^*}{\partial c_i}$ and $\frac{\partial s_i^*}{\partial G}$ can be found by (5.13) and $\frac{dG}{dc_i}$ is in (5.11). One can show that $\frac{\partial s_i^*}{\partial c_i} = \frac{\partial s_j^*}{\partial c_i} + 1$ and $\frac{\partial s_i^*}{\partial G} = \frac{\partial s_j^*}{\partial G}$. Thus, it holds that $\frac{ds_i^*}{dc_i} = \frac{ds_j^*}{dc_i} + 1$. Assume $\frac{\partial s_i^*}{\partial c_i} \leq 0$, then $\frac{\partial s_j^*}{\partial c_i} < 0$ and $\frac{dG}{dc_i} = \sum_{j \in \mathcal{V}} \frac{ds_j^*}{dc_i} < 0$. This contradicts to (5.11). Therefore, $\frac{\partial s_i^*}{\partial c_i} > 0$.

(P4) It holds that

$$\begin{aligned} &\frac{(|\mathcal{V}_a| - 1)(G - \bar{c})}{R + G - \bar{c}} + \sum_{i \in \mathcal{V}_a} \beta_i + 1 - |\mathcal{V}_a| + \frac{R \sum_{i \in \mathcal{V} \setminus \mathcal{V}_a} c_i}{(R + G - \bar{c})^2} \\ &\leq \sum_{i \in \mathcal{V}} \frac{\partial h_i(G)}{\partial G} \leq \frac{(N-1)(G - \bar{c})}{R + G - \bar{c}} + \bar{\beta} + 1 - N \end{aligned} \tag{5.14}$$

where the lower bound can be found from (5.9) and the upper bound can be obtained from (5.8):

$$\sum_{i \in \mathcal{V}} \frac{\partial h_i(G)}{\partial G} \leq -\frac{R(N-1)}{R + G - \bar{c}} + \bar{\beta} \leq \frac{(N-1)(G - \bar{c})}{R + G - \bar{c}} + \bar{\beta} + 1 - N.$$

If $i \in \bar{\mathcal{V}}_a$, then $i \in \mathcal{V}_a$ because it holds that $s_i^* \geq c_i + \frac{(G+R-\bar{c})^2}{R}\Big(\frac{R}{R+G-\bar{c}} + \frac{\partial h_i(G)}{\partial G} - 1\Big) \geq c_i + R\Big(\frac{R}{R+G^U-\bar{c}} + \frac{\partial h_i(G^U)}{\partial G} - 1\Big)$ by equation (5.12). Remind that $\bar{\beta} = N$, and equation (5.14) implies that $\bar{G} \leq G \leq \underline{G}$ because H is a strictly decreasing function. It holds that $\underline{G} \leq G^*$ because $G^* = H^{-1}(1)$ and H^{-1} is also strictly decreasing. Since $\sum_{i \in \mathcal{V}} h_i(G) - G$ is strictly increasing in $G \in [0, G^*]$ and has a maximum at $G = G^*$, we have

$$\sum_{i \in \mathcal{V}} h_i(\bar{G}) - \bar{G} \leq \sum_{i \in \mathcal{V}} U_i(s^*) \leq \sum_{i \in \mathcal{V}} h_i(\underline{G}) - \underline{G}. \tag{5.15}$$

Dividing

$$\max_{s\in\mathcal{S}} \sum_{i\in\mathcal{V}} U_i(s) = \max_{s\in\mathcal{S}} \sum_{i\in\mathcal{V}} h_i(\bar{s}-R) - (\bar{s}-R) = \sum_{i\in\mathcal{V}} h_i(G^*) - G^*$$

by (5.15) yields the desired result.

Now we proceed to prove that PoA > 1 with any $R < \infty$ if $c = \vec{0}$, but $\lim_{R\to\infty} \mathrm{PoA}(R, \vec{0}) = 1$. It can be shown that $\underline{G} = H^{-1}(\frac{(|\bar{\mathcal{V}}_a|-1)\bar{G}}{R+G^U-\bar{c}} + 1) < H^{-1}(1) = G^*$ where $\bar{G} \neq 0$. Therefore, $1 < \frac{\sum_{i\in\mathcal{V}} h_i(G^*)-G^*}{\sum_{i\in\mathcal{V}} h_i(\underline{G})-\underline{G}} \leq \mathrm{PoA}$ with any $R < \infty$. Moreover, as $R \to \infty$, it holds that $\lim_{R\to\infty} \bar{G} = \lim_{R\to\infty} \underline{G} = H^{-1}(1) = G^*$. Therefore, we can conclude that $\lim_{R\to\infty} \mathrm{PoA} = 1$. ■

(P2) shows that there payoff (5.4) does not have discontinuity because $\bar{s}(R, c) > \bar{s}(R, c) - R \geq \bar{c}$. Remind that $\mathrm{PoA}(R, c) = 1$ if and only if $G(R, c) = G^*$. So (P4) indicates that it is impossible to achieve optimality $G(R, c) = G^*$ with a finite reward when perturbations are not allowed; i.e., there is no finite maximizer of problem (5.2). Price of anarchy is identical to Price of stability [141] which represents the ratio between the socially optimal payoff and the aggregate payoff induced by the best Nash equilibrium because there exists a unique Nash equilibrium by (P1). (P3) shows that problem (5.5) is not well-defined neither because $s_i^*(R, c)$ increases unbounded as R increases.

Some properties of Theorem 5.4.1 reduce to those in [60] where perturbations are absent. In particular, (P1) reduces to Proposition 2 of [60] where an unperturbed lottery has a unique Nash equilibrium. (P4) is consistent with Theorem 2 in [60]; i.e., given any $\epsilon > 0$, there exists R such that $\mathrm{PoA}(R, \vec{0}) \leq 1 + \epsilon$. The lower and upper bounds of price of anarchy are newly derived in this chapter and they can be calculated without finding the Nash equilibrium. Additionally, (P2), and (P3) are new and reveal the properties regarding public goods and investment, respectively.

5.5 Convex approximation of high-level social optimum

Problem (5.6) is a bi-level optimization problem. In general, this class of problems is computationally challenging. In particular, papers [142–144] show that

bi-level linear programs are NP-hard. Given the computational hardness, certain relaxations of problem (5.6) are needed in order to find computationally efficient solvers. We will leverage Theorem 5.4.1 to show that the following problem is a convex over-approximation for problem (5.6). We will also show that, under certain mild conditions, the approximation gap is zero.

$$\begin{aligned}
&\min_{(R,c)\in\mathcal{R}\times\mathcal{C}} R + \alpha G^U \\
&\text{s.t. } \bar{c} = G^U,\ R \geq G^U \\
&g(c_1 + R\frac{\partial h_1(G^U)}{\partial G} + R(1-\beta_1), \cdots, c_N + R\frac{\partial h_N(G^U)}{\partial G} + R(1-\beta_N), R, c) \leq 0, \\
&\sum_{i\in\mathcal{V}}(1-\beta_i)(c_i + R\frac{\partial h_i(G^U)}{\partial G} + R(1-\beta_i)) \geq \sum_{i\in\mathcal{V}} h_i(G^*) - G^* - (\sum_{i\in\mathcal{V}} h_i(G^U) - G^U)
\end{aligned} \tag{5.16}$$

The problem (5.16) is convex. The objective function is affine, constraints $\bar{c} = G^U$ and $R \geq G^U$ are also affine. Constraint

$$g(c_1 + R\frac{\partial h_1(G^U)}{\partial G} + R(1-\beta_1), \cdots, c_N + R\frac{\partial h_N(G^U)}{\partial G} + R(1-\beta_N), R, c) \leq 0 \tag{5.17}$$

is convex because a composition of convex function with affine functions preserves convexity where g is a convex function by Assumption 5.3.2 and $c_i + R\frac{\partial h_i(G^U)}{\partial G} + R(1-\beta_i)$ is an affine function. The last constraint is also an affine function in (R, c). Feasible set $\mathcal{R}\times\mathcal{C}$ is convex because $\mathcal{R} = (0,\infty)$, $\mathcal{C}_i = [0, G^U]$ are convex sets and Cartesian products preserve convexity.

Set notations $\mathcal{F}_{(5.6)}$, and $\mathcal{F}_{(5.16)}$ denote the feasible sets of problems (5.6), and (5.16) respectively. Likewise, we define $p^*_{(5.6)}$, and $p^*_{(5.16)}$ as the optimal values of problems (5.6), and (5.16) respectively. The following theorem shows that problem (5.16) is a convex over-approximation of problem (5.6) and the approximation is exact under certain mild conditions.

Theorem 5.5.1 *Under Assumptions 5.2.1, 5.3.1, and 5.3.2, the followings hold:*

- *If $\beta_i = 1$ for $\forall i$, then $\mathcal{F}_{(5.16)} = \mathcal{F}_{(5.6)}$ and $p^*_{(5.16)} = p^*_{(5.6)}$;*
- *If $\beta_i \neq 1$ for some i, then $\mathcal{F}_{(5.16)} \subseteq \mathcal{F}_{(5.6)}$ and $p^*_{(5.16)} \geq p^*_{(5.6)}$. Moreover, if*

$g(s^*(R,c),R,c) \leq 0$ *implies* $\bar{c} = G^U$*, then* $\mathcal{F}_{(5.16)} = \mathcal{F}_{(5.6)}$ *and* $p^*_{(5.16)} = p^*_{(5.6)}$.

PROOF. In the proof, we will drop the dependency of G, s^*, and U_i on R and c. The proofs are divided into three claim statements.

CLAIM 1. $\mathcal{F}_{(5.16)}$ *is a subset of* $\mathcal{F}_{(5.6)}$.

PROOF. Assume that $\mathcal{F}_{(5.16)}$ is non-empty and we pick any $(R,c) \in \mathcal{F}_{(5.16)}$. We will show that such the pair (R,c) satisfies all the constraints in (5.6); i.e., $(R,c) \in \mathcal{F}_{(5.6)}$.

The constraint $\bar{c} = G^U$ implies that $G^U = \bar{c} \leq G \leq G^U$ by (P2). Therefore, it holds that $G = G^U$.

Using $\bar{s}^* - R = G^U = \bar{c}$, the first order condition yields

$$\begin{aligned}\frac{\partial U_i(s^*)}{\partial s_i} &= R\frac{\bar{s}^* - \bar{c} - (s_i^* - c_i)}{(\bar{s}^* - \bar{c})^2} + \frac{\partial h_i(\bar{s}^* - R)}{\partial G} - \beta_i \\ &= 1 - \frac{s_i^* - c_i}{R} + \frac{\partial h_i(G^U)}{\partial G} - \beta_i \leq 0. \end{aligned} \tag{5.18}$$

This equation implies that

$$s_i^* \geq c_i + R\frac{\partial h_i(G^U)}{\partial G} + R(1-\beta_i) > 0 \tag{5.19}$$

because $R \geq G^U > 0$. Since the players are active, equality holds in the first order condition (5.18) as well as (5.19):

$$s_i^* = c_i + R\frac{\partial h_i(G^U)}{\partial G} + R(1-\beta_i). \tag{5.20}$$

Therefore, constraint (5.17) implies

$$g(s^*, R, c) \leq 0. \tag{5.21}$$

By substituting relation (5.20) to the last constraint in (5.16)

$$\begin{aligned}&\sum_{i\in\mathcal{V}}(1-\beta_i)(c_i + R\frac{\partial h_i(G^U)}{\partial G} + R(1-\beta_i)) \\ &= \sum_{i\in\mathcal{V}}(1-\beta_i)s_i^* \geq \sum_{i\in\mathcal{V}} h_i(G^*) - G^* - \sum_{i\in\mathcal{V}} h_i(G^U) + G^U\end{aligned}$$

we have the last constraint in (5.6).

Lastly, the constraint $\bar{c} \leq \min\{G^U, R\}$ holds obviously. Therefore, $(R, c) \in \mathcal{F}_{(5.6)}$. The statement holds because we pick arbitrary $(R, c) \in \mathcal{F}_{(5.16)}$. ■

Claim 1 shows that $\mathcal{F}_{(5.16)} \subseteq \mathcal{F}_{(5.6)}$ for any set of β. The objective function of (5.16) is $\min_{(R,c)} R + \alpha G^U = \min_{(R,c)} R + \alpha\bar{c}$. Therefore, solution $p^*_{(5.16)}$ is an overestimate of $p^*_{(5.6)}$. We now proceed to prove that $\mathcal{F}_{(5.6)} \subseteq \mathcal{F}_{(5.16)}$ if $\beta_i = 1$ and thus $p^*_{(5.6)} = p^*_{(5.16)}$.

Claim 2. *$\mathcal{F}_{(5.6)}$ is a subset of $\mathcal{F}_{(5.16)}$ if $\beta_i = 1$ for $\forall i \in \mathcal{V}$.*

PROOF. Assume that $\mathcal{F}_{(5.6)}$ is non-empty and we pick any $(R, c) \in \mathcal{F}_{(5.6)}$. We will show that the pair satisfies all the constraints in (5.16).

If $\beta_i = 1$ for $\forall i$, constraint

$$\sum_{i\in\mathcal{V}} h_i(G(R,c)) - G(R,c) + \sum_{i\in\mathcal{V}}(1-\beta_i)s_i^* = \sum_{i\in\mathcal{V}} h_i(G(R,c)) - G(R,c)$$
$$\geq \sum_{i\in\mathcal{V}} h_i(G^*) - G^*$$

holds only when $G = G^*$ because G^* is a unique maximizer of the aggregate payoff of unperturbed lottery. Therefore, the relation $G = G^*$ holds.

We now prove $\bar{c} = G^*$ by contradiction. Assume that there exist pair (R, c) such that $G = G^*$ but $\bar{c} \neq G^*$. By (5.14),

$$\frac{(|\mathcal{V}_a| - 1)(G - \bar{c})}{R + G - \bar{c}} + 1 + \frac{R\sum_{i\in\mathcal{V}\setminus\mathcal{V}_a} c_i}{(R + G - \bar{c})^2} \leq \sum_{i\in\mathcal{V}} \frac{\partial h_i(G)}{\partial G}.$$

Since $\sum_{i\in\mathcal{V}} \frac{\partial h_i(G^*)}{\partial G} = 1$, it must hold that $|\mathcal{V}_a| = 1$. First order condition (5.7) for $i \in \mathcal{V}_a$ must hold with equality. However, we have

$$\frac{\partial U_i(s^*)}{\partial s_i} = -R\frac{\bar{c} - c_i}{(G^* + R - \bar{c})^2} + \frac{\partial h_i(G^*)}{\partial G} - 1 < 0$$

which contradicts to the first order condition. Therefore, $\bar{c} = G^*$. Note that if $\bar{c} = G^*$, then $G = G^*$ by (P2).

First order condition with $\bar{c} = G^*$

$$\frac{\partial U_i(s^*)}{\partial s_i} = -\frac{s_i^*}{R} + \frac{c_i}{R} + \frac{\partial h_i(G^*)}{\partial G} \leq 0$$

implies $s_i^* > 0$, which holds for $\forall i \in \mathcal{V}$; i.e., $|\mathcal{V}_a| = N$.

Using the first order condition (5.19), we can derive

$$s_i^* = c_i + R\frac{\partial h_i(G^*)}{\partial G} \tag{5.22}$$

where equality holds because all the players are active. By plugging (5.22) into constraint (5.21), we obtain (5.17).

With $\beta_i = 1$ and $G^U = G^*$, the both sides of the constraint in (5.16)

$$\sum_{i\in\mathcal{V}}(1-\beta_i)(c_i + R\frac{\partial h_i(G^U)}{\partial G} + R(1-\beta_i)) \geq \sum_{i\in\mathcal{V}} h_i(G^*) - G^* - (\sum_{i\in\mathcal{V}} h_i(G^U) - G^U)$$

become zero. Therefore, it holds always. Constraint $R \geq G^*$ is satisfied since $R \geq \bar{c} = G^*$. Therefore, $(R, c) \in \mathcal{F}_{(5.16)}$. The statement holds because we pick arbitrary $(R, c) \in \mathcal{F}_{(5.6)}$. ■

If $\beta_i = 1$ for $\forall i$, $\mathcal{F}_{(5.6)} = \mathcal{F}_{(5.16)}$ by Claim 1 and Claim 2, and the objective functions are equivalent to each other because $\bar{c} = G^*$ for the both feasible sets. Thus it holds that $p^*_{(5.6)} = p^*_{(5.16)}$. Lastly, we proceed to prove that $\mathcal{F}_{(5.6)} \subseteq \mathcal{F}_{(5.16)}$ if $\beta_i \neq 1$ and $g(s^*(R, c), R, c) \leq 0$ implies $\bar{c} = G^U$. This will implies $p^*_{(5.6)} = p^*_{(5.16)}$ because of the equivalence of the objective functions.

Claim 3. *$\mathcal{F}_{(5.6)}$ is a subset of $\mathcal{F}_{(5.16)}$ if $\beta_i \neq 1$ for some i, and $g(s^*(R, c), R, c) \leq 0$ implies $\bar{c} = G^U$.*

PROOF. Assume that $\mathcal{F}_{(5.6)}$ is non-empty and we pick any $(R, c) \in \mathcal{F}_{(5.6)}$. We will show that the pair satisfies all the constraints in (5.16).

First order condition becomes

$$\frac{\partial U_i(s^*)}{\partial s_i} = (1-\beta_i) - \frac{s_i - c_i}{R} + \frac{\partial h_i(G^*)}{\partial G} \leq 0 \tag{5.23}$$

which holds only if $s_i^* > 0$ for $\forall i$; i.e., $|\mathcal{V}_a| = N$. Therefore, equality holds in (5.23) which yields

$$s_i^* = c_i + R\frac{\partial h_i(G^*)}{\partial G} + R(1-\beta_i).$$

By substituting the above relation with $\bar{c} = G^U$ to the constraints, we obtain all the constraints in (5.16). ■

By Claim 1 and Claim 3, $\mathcal{F}_{(5.6)} = \mathcal{F}_{(5.16)}$, if $\beta_i \neq 1$ for some i, and $g(s^*(R, c),$

$R, c) \leq 0$ implies $\bar{c} = G^U$. It also holds that $p^*_{(5.6)} = p^*_{(5.16)}$ because $\bar{c} = G^U$ for the both feasible sets. ■

In Theorem 5.5.1, non-convex optimization problem (5.6) is approximated by (or equivalent to) a convex optimization problem (5.16). In particular, with $\bar{c} = G^U$, we could obtain a constant public good $G(R, c) = G^U$ by (P2) in Theorem 5.4.1, which sequentially results in the replacement of potentially non-concave function $s^*_i(R, c)$ with a linear function (5.20). We show that this condition is a sufficient and necessary condition for the social optimum when $\beta_i = 1$ for $\forall i$, or $g(s^*(R, c), R, c) \leq 0$ implies $\bar{c} = G^U$. On the other hand, it is a sufficient condition when $\beta_i \neq 1$ for some i because a sufficiently large reward R with $\bar{c} \neq G^U$ can induce the optimal aggregate payoff or a greater one as shown in (P3).

5.6 Simulation

In this section, we will apply our perturbed lottery to demand response in the smart grid. Demand response involves a load serving entity (LSE) and a set of end-users. The LSE is the social planner and wants to incentivize the end-users to shift their peak-time demand to off-peak time. The end-users participate the lottery by shifting a portion of their shiftable demands.

Consider a power transmission network described by $(\mathcal{G}, \mathcal{E})$ where $\mathcal{G}$ and $\mathcal{E}$ denote the set of buses and the set of transmission lines, respectively. In particular, $\mathcal{V} \subseteq \mathcal{G}$, and $\mathcal{P} \subseteq \mathcal{G}$ denote the set of load buses with non-zero demand (end-users), and the set of generator buses, respectively. Each line $l \in \mathcal{E}$ has power flow capacity $f^{\max}_l \in \mathbb{R}_{\geq 0}$ and $f^{\max} = [f^{\max}_1, \cdots, f^{\max}_{|\mathcal{E}|}]^T$.

With the perturbed lottery, each end-user has payoff function (5.4) where decision variable s_i denotes shifted demand in monetary value. Heterogeneity parameter β_i denotes the dis-utility incurred by shifting a unit demand. Function h_i represents any impetus from the marginal benefit; e.g., utility discount, additional rewards, public good made by the LSE. The LSE solves problem (5.6), in which convex constraints represent three physical constraints; i.e., the end-users cannot shift more than the demand, and the total adjusted demand after shifting cannot exceed the total power generation, and the line capacities are enforced:

$$L - s^*(R, c) \geq \vec{0}\,,\ \sum_{i \in \mathcal{V}} (L_i - s^*_i(R, c)) \leq \sum_{j \in \mathcal{P}} P_j,$$

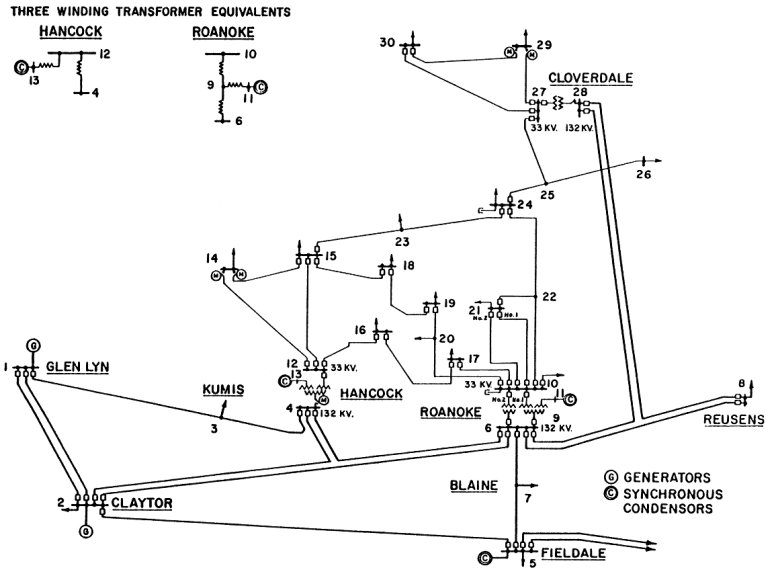

Figure 5.1: IEEE 30-bus test system [3].

$$-f^{\max} \leq H_p P - H_l(L - s^*(R,c)) \leq f^{\max} \tag{5.24}$$

where $L \in \mathbb{R}^{|\mathcal{V}|}_{\geq 0}$ and $P \in \mathbb{R}^{|\mathcal{P}|}_{\geq 0}$ denote power demand and power generation, respectively. Matrix $H \in [-1,1]^{|\mathcal{E}|\times|\mathcal{G}|}$ is the injection shift factor matrix where (a,b) entry of H represents the active power change on line a with respect to change in power injection at bus b. Matrices $H_l \in [-1,1]^{|\mathcal{E}|\times|\mathcal{V}|}$ and $H_p \in [-1,1]^{|\mathcal{E}|\times|\mathcal{P}|}$ are the collections of columns $i \in \mathcal{V}$ and $i \in \mathcal{P}$ of H, respectively. Since L, P, $f^{\max}$ are constants at the given time, constraints (5.24) are convex and thus satisfy Assumption 5.3.2.

We conduct case studies using IEEE 30-bus test system shown in Figure 5.1 where $|\mathcal{P}| = 6$, $|\mathcal{V}| = 20$, and $|\mathcal{E}| = 41$. The system parameters are obtained from MATPOWER [145]. Money/power exchange rate $\$0.1/kWh$ is applied and 1 hour time frame is considered; e.g., the generator at bus 1 generates $23.54MW \times 1h \times \$0.1/kWh = \$2354$. We intentionally increase the power demand of each load bus by 30% without changing power generations, so that demand shifts are inevitable.

We choose $h_i(\bar{s} - R) = (100 + i)\log(\bar{s} - R + 1)$ for bus $i \in \mathcal{V}$; e.g., bus $30 \in \mathcal{V}$ has $h_{30}(\bar{s} - R) = 130\log(\bar{s} - R + 1)$. One can see that function h_i satisfies Assumption 5.2.1. The logarithmic model of provision of public good h_i is based on Cobb-Douglas utility function [146]. Recent papers [147–149] use such function

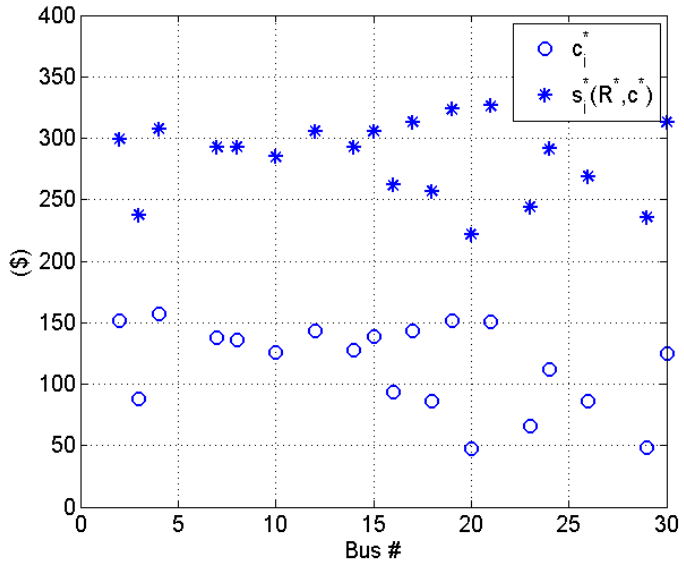

Figure 5.2: ($\beta_i = 1$) Optimal solution c_i^* with $R^* = \$3358$ and the corresponding Nash equilibrium s_i^*.

to express the benefit from a public good. We choose $\alpha = 1$.

The socially optimal public good $G^* = \$2317$ of unperturbed lottery is calculated by (5.3). The socially optimal payoff is obtained by $\sum_{i\in\mathcal{V}} h_i(G^*) - G^* = \$7142$.

Case study 1 ($\beta_i = 1$). We solve problem (5.6) by CVX [150], and generate optimal value \$5675 with solution (R^*, c^*) presented in Figure 5.2. The figure also presents the induced Nash equilibrium of the optimal lottery game. The aggregate payoff induces the socially optimal public good $\bar{s}^*(R^*, c^*) - R^* = \$5675 - \$3358 = \$2317 = G^*$, and the socially optimal payoff $\sum_{i\in\mathcal{V}} h_i(G^*) - G^* = \$15644$. Convex program (5.16) generates a large reward $R^* = \$3358 > \$2317 = G^*$ to satisfy the physical constraints. Note that $\bar{c} = G^*$ with $R \geq G^*$ is a sufficient and necessary condition for the optimality, according to Theorem 5.5.1. By Theorem 5.5.1, the solution is identical to that of problem (5.6) and satisfies all the physical constraints described in (5.24). The left hand side of Figure 5.3 visualizes that the first constraint is satisfied where the shifted demand never exceeds the power demand. The second constraint is also satisfied because $\sum_{j\in\mathcal{P}} P_j = \sum_{i\in\mathcal{V}}(L_i - s_i^*(R^*, c^*)) = \$18921$. The right hand side of Figure 5.3 shows that power flow at each transmission line never exceeds its capacity.

Case study 2 ($\beta_i \neq 1$). We choose $\beta_i = 0.99$ for the first 10 end-users and $\beta_i =$

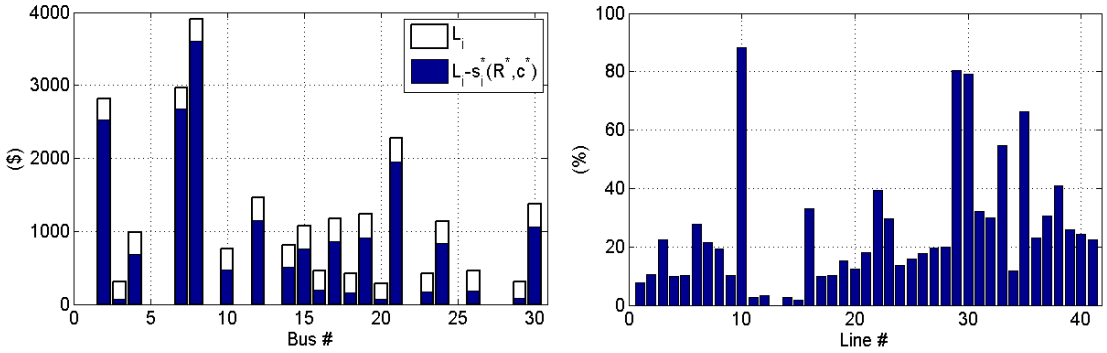

Figure 5.3: ($\beta_i = 1$) Power demand and adjusted demand after shifting, and percentage of power flow used in each line.

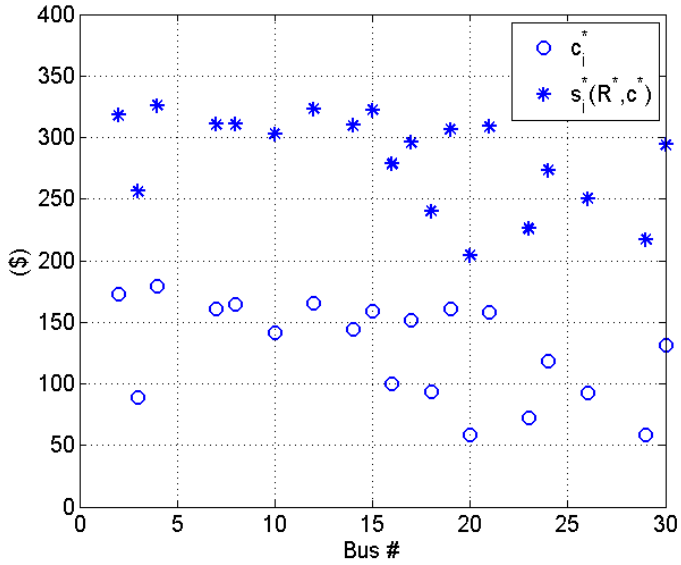

Figure 5.4: ($\beta_i \neq 1$) Optimal solution c_i^* with $R^* = \$3100.4$ and the corresponding Nash equilibrium.

1 for the remaining end-users. Value $G^U = \$2574.6$ is found by $\sum_{i\in\mathcal{V}} \frac{\partial h_i(G^U)}{\partial G} = 0.9$. The optimal value of problem (5.6) is found by \$5675 and its solution is presented in Figure 5.4. The Nash equilibrium is also presented in the same figure, and it induces the public good $\bar{s}^*(R^*, c^*) - R^* = \$5675 - \$3100.4 = \$2574.6 = G^U$, as designed. The aggregate payoff $\sum_{i\in\mathcal{V}} h_i(G^U) - G^U + \sum_{i\in\mathcal{V}}(1-\beta_i)s_i^*(R^*, c^*) = \$15631 + \$3058.5$ is greater than that of the socially optimal payoff \$15644 as desired. Since $g(s^*(R, c), R, c) \leq 0$ does not imply $\bar{c} = G^U$, the solution might be sub-optimal, but all the physical constraints (5.24) are satisfied. As Case 1,

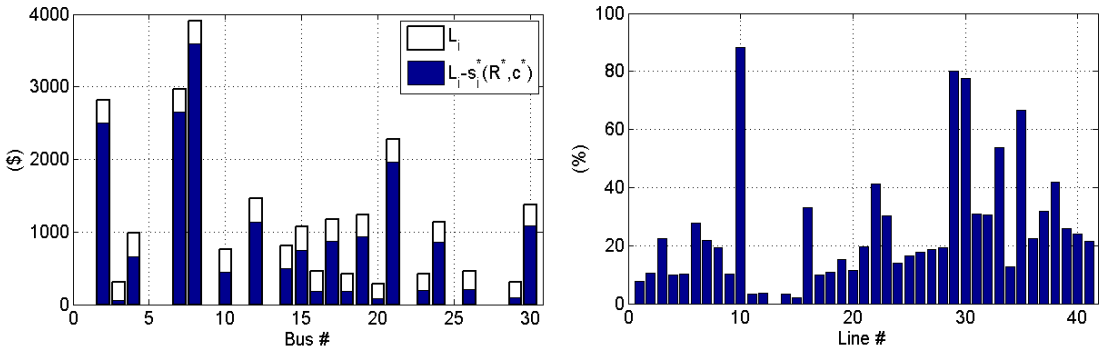

Figure 5.5: ($\beta_i \neq 1$) Power demand and adjusted demand after shifting and percentage of power flow used in each line.

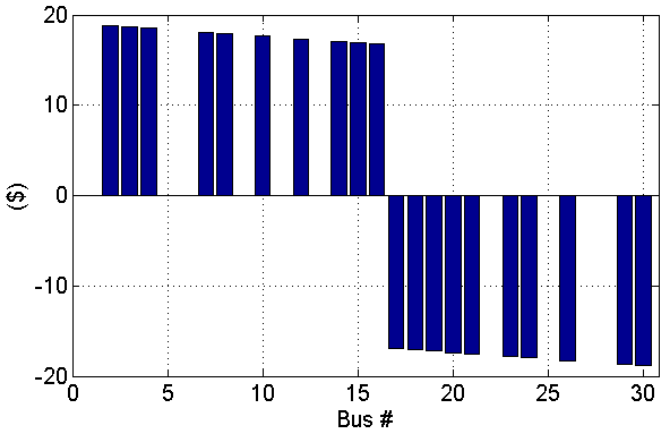

Figure 5.6: Difference of Nash equilibriums of Case 1 and Case 2; s_i^*(Case2) − s_i^*(Case1).

Figure 5.5 shows that the shifted demand does not exceed the power demand, and the amount of power flow in line is below its limit. Lastly, the total adjusted demand after shifting is $\sum_{j\in\mathcal{P}} P_j = \sum_{i\in\mathcal{V}}(L_i - s_i^*(R^*, c^*)) = \$18921$. The Nash equilibriums of Case 1 and 2 are different although the aggregate shifted demand is identical. It is mainly because they have heterogeneous payoff functions in each case. In particular, $\beta_i < 1$ implies a willingness to shift more demand; i.e., the first 10 end-users in Case 2 are more likely to shift demand than the others. Thus, even with a smaller reward than that in Case 1, the LSE is able to induce the same amount of shifted demand. The difference of Nash equilibriums is visualized in Figure 5.6 where the first 10 users shift more demands in Case 2 than in Case 1.

5.7 Conclusion

In this chapter, we study an optimal bi-level lottery design problem where a social planner aims to achieve the social optimum through least reward and perturbations. We approximate the problem via a convex relaxation and identify mild sufficient conditions under which the approximation is exact. The results are verified via a case study on demand response in the smart grid.

5.8 Appendix

Proof of Lemma 5.3.1. We prove the statement by construction. We will show that a pair (R, c) such that $c_i = \frac{G^U}{N}$, $R = \max\{R_L + 1, G^U, \bar{R}\}$, satisfies $\sum_{i\in\mathcal{V}} h_i(G) - G + \sum_{i\in\mathcal{V}}(1-\beta_i)s_i^* \geq \sum_{i\in\mathcal{V}} h_i(G^*) - G^*$ where $\bar{R}$ will be defined later. By (P2), $\bar{c} = G = G^U$. If $\beta_i = 1$ for $\forall i$, then $G^U = G^*$ and thus the pair satisfies the constraint with $\bar{R} = 0$.

Assume $\beta_i \neq 1$ for some i. It holds that $s_i^* \geq c_i + R\Big(\frac{R}{R+G^U-\bar{c}} + \frac{\partial h_i(G^U)}{\partial G} - \beta_i\Big)$ by (P3) where the lower bound is a strictly increasing function in R without bound. Therefore, there always exists sufficiently large $\bar{R}$ such that

$$\sum_{i\in\mathcal{V}}(1-\beta_i)s_i^* \geq \sum_{i\in\mathcal{V}} h_i(G^*) - G^* - \sum_{i\in\mathcal{V}} h_i(G^U) - G^U$$

which satisfies the constraint. ■

Chapter 6
Frequency control

6.1 Introduction

The traditional power grid is modernized into the smart grid. The wide deployment of advanced information and communications technologies facilitates real-time pricing and demand response. In addition, centralized generating facilities are giving way to small distributed energy resources; e.g., photovoltaic systems, fuel cells, storage and electric vehicles. Moreover, renewable energy; e.g., wind, solar and wave energy, has been increasingly adopted due to its cleanness and profitability.

While the integration allows flexible and efficient management of the grid, it leaves uncertainties on the smart grid as well. In particular, uncertain renewable generation as well as uncontrollable (unknown) loads can be seen as external disturbances to the smart grid. As discussed in Chapter 1, existing papers focus on control methodologies for disturbance attenuation where the impact of disturbances reduces but does not completely disappear. In disturbance attenuation, the impact increases as the disturbance increases. Since the smart grid integrates numerous uncertain components, it is imperative to study disturbance rejection where the impact of disturbance is completely removed, regardless of the size of disturbance. We, in this chapter, study frequency control in the presence of uncertain disturbances, and design disturbance rejecting distributed controllers.

Chapter organization. We consider frequency control problems in the presence of uncertain net loads. Power system model with synchronous generator and loads is introduced in Section 6.2.1. In the same section, robust frequency con-

trol problem and robust adaptive frequency control problem are illustrated. We propose distributed controllers to address the problems in Sections 6.3 and 6.4. Stability of the designed controllers are analyzed and their proofs are presented in Section 6.5. Lastly, numerical simulations demonstrate the performance in Section 6.6.

Notations. Denote $\|x\|_{[t_1,t_2]} \triangleq \sup_{t_1 \le t \le t_2} \|x(t)\|$. Let $|\mathcal{S}|$ be the cardinality of a set $\mathcal{S}$. Matrix I_n denotes the $n \times n$ identity matrix. Let $\text{diag}(A_1, \cdots, A_n)$ denote a block matrix having A_1 to A_n as main diagonal blocks. For $v \in \mathbb{R}^n$, $sgn(v) \in \{-1, 0, 1\}^n$ is a sign function, where $sgn(v_i) = -1$ if $v_i < 0$, $sgn(v_i) = 0$ if $v_i = 0$, and $sgn(v_i) = 1$ if $v_i > 0$. Norm $\|\cdot\|_F$ denotes Frobenius norm.

6.2 Problem formulation

In this section, we present a model of power systems, and frequency control problems.

6.2.1 System model

Table 6.1 summarizes the notations used in the model. We use Δ to represent deviations from nominal values; e.g., $\Delta w(t) = w(t) - w^*$, where w^* is the nominal value of $w(t)$.

Power network model The power network is described by the undirected graph $(\mathcal{V}, \mathcal{E})$ where $\mathcal{V} \triangleq \{1, \cdots, N\}$ denotes the set of buses and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ denotes the set of transmission lines between the buses. The set $\mathcal{N}_i$ denotes the set of neighboring buses of $i \in \mathcal{V}$; i.e., $\mathcal{N}_i \triangleq \{j \in \mathcal{V} \setminus \{i\} | (i,j) \in \mathcal{E}\}$. Each bus is either a generator bus $i \in \mathcal{G}$, or a load bus $i \in \mathcal{L}$ where $\mathcal{G}$ and $\mathcal{L}$ denote the sets of corresponding buses, respectively. Each bus i is associated with a local control authority.

Load model An electrical load can be divided to a controllable load and an uncontrollable load [85, 151, 152]. Controllable load $\Delta P_{C_i}(t)$ is governed by the demand response [85]:

$$\Delta \dot{P}_{C_i}(t) = b_i + c_i \Delta P_{C_i}(t) - \Delta \lambda_i(t) \tag{6.1}$$

where $b_i + c_i \Delta P_{C_i}(t)$ is marginal benefit with $c_i < 0$ and real-time electricity price

Table 6.1: System variables and parameters.

System variables			
w	angular frequency	θ	phase angle
P_M	mechanical power	P_{ij}	power flow
P_v	steam valve position	P_{ref}	reference power
P_C	controllable load	P_L	net load
System parameters			
D	damping constant	m	angular momentum
T_{CH}	charging time const.	t_{ij}	tie-line stiffness
T_G	governor time const.	R	feedback loop gain
K_m	turbine gain	K_e	governor gain

$\Delta\lambda_i(t)$ is used as the input for $i \in \mathcal{L}$.

For $i \in \mathcal{L}$, net load $\Delta P_{L_i}(t)$ represents the difference between the uncontrollable load and renewable generation. For $i \in \mathcal{G}$, net load $\Delta P_{L_i}(t)$ represents renewable generation. Notice that uncontrollable loads and renewable generation are hard to predict. We regard net loads $\Delta P_{L_i}(t)$ as external disturbances to the power system. According to the spectral decompositions of wind generation [153, 154] and load pattern [155], we approximate each net load as the sum of a finite number of distinct sinusoidal functions as in [65]. In addition, any periodic function can be represented by a Fourier series. If a function is continuous, absolutely integrable and its derivative is absolutely integrable, then its Fourier series converges uniformly to the function (Theorem on p.86 in [156]). As output regulation [157,158], the following marginally stable exosystem is used to generate $\Delta P_{L_i}(t)$:

$$\dot{\chi}_i(t) = \Phi_i(\rho_i)\chi_i(t),\ \Delta P_{L_i}(t) = \Psi_i\chi_i(t) \tag{6.2}$$

where $\chi_i(t) = [\chi_{i,1}(t), \dot{\chi}_{i,1}(t), \cdots, \chi_{i,\ell_i}(t), \dot{\chi}_{i,\ell_i}(t)]^T \in \mathbb{R}^{2\ell_i}$,

$$\Phi_i(\rho_i) \triangleq \text{diag}(\Phi_{i,1}, \cdots, \Phi_{i,\ell_i}),\ \Phi_{i,l} \triangleq \begin{bmatrix} 0 & 1 \\ -(\rho_{i,l})^2 & 0 \end{bmatrix}.$$

Each state $\chi_{i,l}(t)$ is a sinusoidal function with frequency $\rho_{i,l}$. The output $\Delta P_{L_i}(t)$ is then a linear combination of sinusoidal functions with frequencies $\rho_i = \{\rho_{i,1}, \cdots, \rho_{i,\ell_i}\}$.

Assumption 6.2.1 *The pair* $(\Psi_i, \Phi_i(\rho_i))$ *is observable.*

Dynamic model of the generator buses Consider the synchronous power

generator from [80]:

$$\begin{aligned}
\Delta\dot{\theta}_i(t) &= \Delta w_i(t) \\
\Delta\dot{w}_i(t) &= -\frac{1}{m_i}\Big((D_{G_i}+D_{L_i})\Delta w_i(t) + \sum_{j\in\mathcal{N}_i}\Delta P_{ij}(t) + \Delta P_{L_i}(t) - \Delta P_{M_i}(t)\Big) \\
\Delta\dot{P}_{M_i}(t) &= -\frac{1}{T_{CH_i}}\Big(\Delta P_{M_i}(t) - K_{m_i}\Delta P_{v_i}(t)\Big) \\
\Delta\dot{P}_{v_i}(t) &= -\frac{1}{T_{G_i}}\Big(\Delta P_{v_i}(t) + \frac{K_{e_i}}{R_i}\Delta w_i(t) - \Delta P_{ref_i}(t)\Big)
\end{aligned} \tag{6.3}$$

for $i \in \mathcal{G}$, where D_{G_i} is mechanical damping constant, and D_{L_i} is load damping constant corresponding to net load (renewable generation). Power flow $\Delta P_{ij}(t)$ is described by $\Delta P_{ij}(t) = t_{ij}(\Delta\theta_i(t) - \Delta\theta_j(t))$. The first equation in (6.3) indicates the evolution of phase angle $\Delta\theta_i(t)$. The second equation is referred to as swing dynamics, indicating frequency fluctuations due to power imbalances. The third and fourth equations represent turbine governor dynamics with reference input $\Delta P_{ref_i}(t)$.

Dynamic model of the load buses A load bus $i \in \mathcal{L}$ can be modeled by the following phase angle dynamics and swing dynamics [104], where m_i is the effective moment of a postulated load model:

$$\begin{aligned}
\Delta\dot{\theta}_i(t) &= \Delta w_i(t) \\
\Delta\dot{w}_i(t) &= -\frac{1}{m_i}\Big(D_{L_i}\Delta w_i(t) + \sum_{j\in\mathcal{N}_i}\Delta P_{ij}(t) + \Delta P_{C_i} + \Delta P_{L_i}(t)\Big)
\end{aligned} \tag{6.4}$$

where D_{L_i} is load damping constant corresponding to net load which is the difference between the uncontrollable load and renewable generation.

Outputs and inputs Control authority i can access $\Delta y_i(t) = [\Delta w_i(t), \Delta P_{M_i}(t), \Delta P_{v_i}(t), \Delta P_{\mathcal{N}_i}$ for $i \in \mathcal{G}$, and $\Delta y_i(t) = [\Delta w_i(t), \Delta P_{C_i}(t), \Delta P_{\mathcal{N}_i}(t)]^T$ for $i \in \mathcal{L}$, where $\Delta P_{\mathcal{N}_i}(t) \triangleq \sum_{j\in\mathcal{N}_i}\Delta P_{ij}(t)$. Local inputs are $u_i(t) = \Delta P_{ref_i}(t)$ for $i \in \mathcal{G}$, and $u_i(t) = \Delta\lambda_i(t)$ for $i \in \mathcal{L}$.

6.2.2 Frequency control problems

In this chapter, we investigate the frequency control; i.e., controlling $\Delta w_i(t)$ to zero, and discuss two cases where the frequencies ρ_i of net loads are known or

unknown. We drop Δ in the rest of the chapter for notational simplicity. Also, we use $D_i = D_{G_i} + D_{L_i}$ for $i \in \mathcal{G}$ and $D_i = D_{L_i}$ for $i \in \mathcal{L}$.

Case 1: Robust frequency control. To stabilize the frequencies, each generator bus $i \in \mathcal{G}$ aims to approach the following manifolds:

$$\begin{aligned} w_i^* = 0,\ P_{M_i}^*(t) &= P_{L_i}(t) + P_{\mathcal{N}_i}(t), \\ P_{v_i}^*(t) &= \frac{T_{CH_i}}{K_{m_i}} \dot{P}_{M_i}^*(t) + \frac{1}{K_{m_i}} P_{M_i}^*(t) \\ P_{ref_i}^*(t) &= T_{G_i} \dot{P}_{v_i}^*(t) + P_{v_i}^*(t) \end{aligned} \tag{6.5}$$

which can be easily derived from system (6.3). Similarly, each load bus $i \in \mathcal{L}$ is expected to stay on the following manifolds:

$$\begin{aligned} w_i^* = 0,\ P_{C_i}^*(t) &= -P_{L_i}(t) - P_{\mathcal{N}_i}(t), \\ \lambda_i^*(t) &= b_i + c_i P_{C_i}^*(t) - \dot{P}_{C_i}^*(t) \end{aligned} \tag{6.6}$$

which can be derived from systems (6.1) and (6.4). Superscript $*$ denotes the manifold; e.g., $\lambda_i^*(t)$ denotes the manifold of $\Delta\lambda_i(t)$. We desire to design a distributed controller which steers the system states and inputs to their manifolds (6.5) and (6.6). In this case, control authority i knows ρ_i, Ψ_i and $\Phi_i(\rho_i)$ but is unaware of initial state $\chi_i(0)$ and state $\chi_i(t)$ of exosystem (6.2). That is, control authority i knows the number of sinusoidal signals, and their frequencies, but not their phase shifts and magnitudes.

Assumption 6.2.2 *All the frequencies $\rho_{i,1}, \cdots, \rho_{i,\ell_i}$ in* (6.2) *of net load $P_{L_i}(t)$ are known to local control authority i.*

Case 2: Robust adaptive frequency control. There are a couple of distinctions from *Case 1.* First, control authority i is unaware of frequencies ρ_i and Assumption 6.2.2 is weakened into the following one:

Assumption 6.2.3 *Control authority i knows the value ℓ_i and an upper bound $\rho_{\max} \geq \max_{i,l} \rho_{i,l}$.*

Secondly, we use the following simplified synchronous generator model [80]:

$$\dot{\theta}_i(t) = w_i(t)$$

$$\dot{w}_i(t) = -\frac{1}{m_i}\Big(D_i w_i(t) + \sum_{j\in\mathcal{N}_i} P_{ij}(t) + P_{L_i}(t) - P_{M_i}(t)\Big) \tag{6.7}$$

and simplified demand response model; i.e., local control authority i controls $P_{C_i}(t)$ directly. Remark 6.4.1 in Section 6.4.3 discusses why the simplified models are needed for Case 2. For this case, the corresponding manifolds are

$$\begin{aligned} &w_i^* = 0,\ P_{M_i}^*(t) = P_{L_i}(t) + P_{\mathcal{N}_i}(t),\ i \in \mathcal{G} \\ &w_i^* = 0,\ P_{C_i}^*(t) = -P_{L_i}(t) - P_{\mathcal{N}_i}(t),\ i \in \mathcal{L}. \end{aligned} \tag{6.8}$$

6.3 Controller synthesis for robust frequency control

In this section, we present a solution of the *robust frequency control* described in Section 6.2.2.

6.3.1 Local internal models

Net loads $P_{L_i}(t)$ cannot be measured and thus manifolds (6.5) and (6.6) cannot be used for feedback control. We adopt the methodology of internal models to tackle this challenge [158,159]. Recall that $P_{L_i}(t)$ is the output of exosystem (6.2). Hence, for $i \in \mathcal{G}$, the second equation in (6.5) can be written as:

$$P_{M_i}^*(t) - P_{\mathcal{N}_i}(t) = P_{L_i}(t) = \Psi_i \chi_i(t). \tag{6.9}$$

Under Assumption 6.2.1, for any controllable pair (M_i, N_i) with $M_i \in \mathbb{R}^{2\ell_i \times 2\ell_i}$ being Hurwitz and $N_i \in \mathbb{R}^{2\ell_i}$, there exists a non-singular matrix $T_i(\rho_i) \in \mathbb{R}^{2\ell_i \times 2\ell_i}$ as the unique solution of the following Sylvester equation [160]:

$$T_i(\rho_i)\Phi_i(\rho_i) - M_i T_i(\rho_i) = N_i \Psi_i. \tag{6.10}$$

With $\vartheta_i(t) \triangleq T_i(\rho_i)\chi_i(t)$, (6.9) becomes

$$P_{M_i}^*(t) - P_{\mathcal{N}_i}(t) = \Psi_i \chi_i(t) = \Psi_i T_i^{-1}(\rho_i)\vartheta_i(t).$$

Now consider a local internal model candidate:

$$\dot{\eta}_i(t) = M_i\eta_i(t) + N_i(P_{M_i}(t) - P_{\mathcal{N}_i}(t)) \tag{6.11}$$

where $\eta_i(t) \in \mathbb{R}^{2\ell_i}$. Internal model (6.11) behaves as an estimator and its states $\eta_i(t)$ are expected to asymptotically track unmeasurable exosystem states $\vartheta_i(t)$. The manifolds of $\eta_i(t)$ are $\eta_i^*(t) = \vartheta_i(t)$ in this case. It is expected to stabilize the dynamics of error $\eta_i(t) - \vartheta_i(t)$. According to the certainty equivalence principle [122], internal model states $\eta_i(t)$ are used to replace $\vartheta_i(t)$ in manifolds (6.5) and then in feedback control.

For load bus $i \in \mathcal{L}$, we derive a similar internal model candidate by replacing $P_{M_i}(t)$ with $-P_{C_i}(t)$:

$$\dot{\eta}_i(t) = M_i\eta_i(t) - N_i(P_{C_i}(t) + P_{\mathcal{N}_i}(t)). \tag{6.12}$$

For notional simplicity, we will use the augmented states $x_i(t) = [x_{i,1}(t), x_{i,2}(t), x_{i,3}(t), x_{i,4}^T(t)]^T = [w_i(t), P_{M_i}(t), P_{v_i}(t), \eta_i^T(t)]^T$ and manifolds $x_i^*(P_{L_i}(t), t) = [x_{i,1}^*(t), x_{i,2}^*(t), x_{i,3}^*(t), (x_{i,4}^*(t))^T]^T = [w_i^*(t), P_{M_i}^*(P_{L_i}(t), t),\ P_{v_i}^*(P_{L_i}(t), t),\ \vartheta_i^T(t)]^T$ for $i \in \mathcal{G}$ and use the augmented states $x_i(t) = [x_{i,1}(t), x_{i,2}(t), x_{i,4}^T(t)]^T = [w_i(t), P_{C_i}(t), \eta_i^T(t)]^T$ and manifolds $x_i^*(P_{L_i}(t), t) = [x_{i,1}^*(t), x_{i,2}^*(t), (x_{i,4}^*(t))^T]^T = [w_i^*(t), P_{C_i}^*(P_{L_i}(t), t), \vartheta_i^T(t)]^T$ for $i \in \mathcal{L}$, where the dependency of x_i^* on $P_{L_i}(t)$ is emphasized.

6.3.2 Controller design

We first conduct a coordinate transformation to convert the frequency control problem into a global stabilization problem of the error dynamics with respect to manifolds (6.5) and (6.6). We make use of its unique lower triangular structure and apply a backstepping technique [161] to stabilize the error dynamics from the outer state to the inner state progressively.

Since internal model states $\eta_i(t)$ are expected to track $\vartheta_i(t)$ asymptotically for $\forall i \in \mathcal{V}$, the estimation errors $\|\Psi_i T_i^{-1}(\rho_i)\eta_i(t) - \Psi_i T_i^{-1}(\rho_i)\vartheta_i(t)\|$ are expected to diminish. By the certainty equivalent principle, we use the known term $\Psi_i T_i^{-1}(\rho_i)\eta_i(t)$ to replace unknown $P_{L_i}(t) = \Psi_i T_i^{-1}(\rho_i)\vartheta_i(t)$ when constructing the error dynamics.

Let us defined the tracking errors as follows:

$$\tilde{x}_i(t) \triangleq x_i(t) - x_i^*(\Psi_i T_i^{-1}(\rho_i)\eta_i(t), t) \tag{6.13}$$

for $\forall i \in \mathcal{V}$. Error dynamics for $i \in \mathcal{G}$ become

$$\begin{aligned}
\dot{\tilde{x}}_{i,1}(t) &= -\frac{1}{m_i}(D_i\tilde{x}_{i,1}(t) - \Psi_i T_i^{-1}(\rho_i)\tilde{x}_{i,4}(t) - \tilde{x}_{i,2}(t)) \\
\dot{\tilde{x}}_{i,2}(t) &= -\frac{1}{T_{CH_i}}(\tilde{x}_{i,2}(t) - K_{m_i}\tilde{x}_{i,3}(t)) \\
\dot{\tilde{x}}_{i,3}(t) &= -\frac{1}{T_{G_i}}(\tilde{x}_{i,3}(t) + \frac{K_{e_i}}{R_i}\tilde{x}_{i,1}(t) - \tilde{P}_{ref_i}(t)) \\
&\quad + \sum_{j\in\mathcal{N}_i} t_{ij}(\Psi_i T_i^{-1}(\rho_j)\tilde{x}_{i,4}(t) - \Psi_j T_j^{-1}(\rho_j)\tilde{x}_{j,4}(t)) \\
\dot{\tilde{x}}_{i,4}(t) &= \Phi_i(\rho_i)\tilde{x}_{i,4}(t) + N_i\tilde{x}_{i,3}(t)
\end{aligned} \tag{6.14}$$

where $\tilde{P}_{ref_i}(t) \triangleq P_{ref_i}(t) - P^*_{ref_i}(\Psi_i T_i^{-1}(\rho_i)x_{i,4}(t), t)$. All the eigenvalues of $\Phi_i(\rho_i)$ are on the imaginary axis. Coordinate transformation $\hat{x}_{i,4}(t) \triangleq \tilde{x}_{i,4}(t) - \hat{x}^*_{i,4}(t)$ leads to

$$\dot{\hat{x}}_{i,4}(t) = M_i\hat{x}_{i,4}(t) + (m_i M_i + D_i I_{2\ell_i})N_i\tilde{x}_{i,1}(t)$$

where $\hat{x}^*_{i,4}(t) = m_i N_i x_{i,1}(t)$. Since matrix M_i is Hurwitz, the subsystem $\hat{x}_{i,4}(t)$ is input-to-state stable (ISS) regarding $\tilde{x}_{i,1}(t)$ as an external input.

Consider subsystem $\tilde{x}_{i,l-1}(t)$ in (6.14) for $l = 2, 3$ and regard $\tilde{x}_{i,l}(t)$ as an external input. Tracking error $\tilde{x}_{i,l}(t)$ is designed to stabilize $\tilde{x}_{i,l-1}(t)$; i.e., the manifold $\hat{x}^*_{i,l}(t)$ of $\tilde{x}_{i,l}(t)$ cancels all the measurable terms in the dynamics of $\tilde{x}_{i,l-1}(t)$ and stabilizes it via $-k_{i,l-1}\tilde{x}_{i,l-1}(t)$. Apply the same idea to $i \in \mathcal{L}$, then we have

$$\hat{x}_i(t) \triangleq \tilde{x}_i(t) - \hat{x}^*_i(t) \tag{6.15}$$

and inputs

$$\begin{aligned}
P_{ref_i}(t) &= P^*_{ref_i}(\Psi_i T_i^{-1}(\rho_i)x_{i,4}(t), t) + (\frac{K_{e_i}}{R_i} + \frac{T_{CH_i}}{K_{m_i}}T_{G_i}(e_i^* + k_{i,1})(e_i^* + k_{i,2}) \\
&\times (D_i - m_i\Psi_i T_i^{-1}(\rho_i)N_i))x_{i,1}(t) - \frac{T_{CH_i}}{K_{m_i}}T_{G_i}(e_i^* + k_{i,1})(e_i^* + k_{i,2})(x_{i,2}(t) - x^*_{i,2}(t))
\end{aligned}$$

$$
\begin{aligned}
&+ T_{G_i}(\frac{1}{K_{m_i}} - \frac{T_{CH_i}}{K_{m_i}}(e_i^* + k_{i,1} + k_{i,2}))(\dot{x}_{i,2}(t) - \dot{x}_{i,2}^*(t)) + (x_{i,3}(t) - x_{i,3}^*(t)) \\
&- k_{i,3} T_{G_i} \hat{x}_{i,3}(t) \\
&\lambda_i(t) = \lambda_i^*(\Psi_i T_i^{-1}(\rho_i) x_{i,4}(t), t) + m_i(e_i^* + k_{i,1})(c_i + k_{i,1}) x_{i,1}(t) \\
&+ (c_i + e_i^* + k_{i,1} + k_{i,2}) \hat{x}_{i,2}(t)
\end{aligned} \tag{6.16}
$$

where $e_i^* \triangleq \Psi_i T_i^{-1}(\rho_i) N_i - \frac{D_i}{m_i}$, $\hat{x}_{i,1}^*(t) = 0$, $\hat{x}_{i,2}^*(t) = -m_i(e_i^* + k_{i,1}) x_{i,1}(t)$, and $\hat{x}_{i,3}^*(t) = -m_i \frac{T_{CH_i}}{K_{m_i}}(e_i^* + k_{i,1})(e_i^* + k_{i,2}) x_{i,1}(t) - (\frac{1}{K_{m_i}} - \frac{T_{CH_i}}{K_{m_i}}(e_i^* + k_{i,1} + k_{i,2}))(x_{i,2}(t) - x_{i,2}^*(t))$. Through transformations (6.13), (6.15) and input (6.16), the augmented system, including (6.1), (6.3), (6.4), (6.11) and (6.12), becomes

$$
\dot{\hat{x}}_i(t) = A_i \hat{x}_i(t) + \sum_{j \in \mathcal{N}_i} B_{ij} \hat{x}_{j,4}(t) \tag{6.17}
$$

where

$$
\begin{aligned}
A_i &= \begin{bmatrix} -k_{i,1} & 1/m_i & 0 & \frac{\Psi_i T_i^{-1}(\rho_i)}{m_i} \\ 0 & -k_{i,2} & K_{m_i}/T_{CH_i} & A_i(2,4) \\ 0 & 0 & -k_{i,3} & A_i(3,4) \\ A_i(4,1) & \mathbf{0}_{2\ell_i \times 1} & \mathbf{0}_{2\ell_i \times 1} & M_i \end{bmatrix}, i \in \mathcal{G} \\
A_i &= \begin{bmatrix} -k_{i,1} & 1/m_i & \Psi_i T_i^{-1}(\rho_i)/m_i \\ 0 & -k_{i,2} & \bar{A}_i(2,4) \\ A_i(4,1) & \mathbf{0}_{2\ell_i \times 1} & M_i \end{bmatrix}, i \in \mathcal{L} \\
A_i(2,4) &= -(e_i^* + k_{i,1}) \Psi_i T_i^{-1}(\rho_i), \\
A_i(4,1) &= (m_i M_i + D_i I_{2\ell_i}) N_i, \\
A_i(3,4) &= -\frac{1}{K_{m_i}}(T_{CH_i}(e_i^* + k_{i,1})(e_i^* + k_{i,2}) + \sum_{j \in \mathcal{N}_i} t_{ij}) \Psi_i T_i^{-1}(\rho_i), \\
B_{ij} &= [\mathbf{0}_{1 \times 2\ell_i}^T, \mathbf{0}_{1 \times 2\ell_i}^T, -\frac{t_{ij}}{K_{m_i}}(\Psi_j T_j^{-1}(\rho_j))^T, \mathbf{0}_{2\ell_i \times 2\ell_i}^T]^T, \text{ for } i \in \mathcal{G} \\
B_{ij} &= \mathbf{0}_{(2\ell_i+1) \times 1}, \text{ for } i \in \mathcal{L}.
\end{aligned} \tag{6.18}
$$

By the backstepping technique, the k_i submatrix in (6.18) is an upper-triangular Hurwitz matrix and M_i is Hurwitz. This property is crucial for the stability of system (6.17).

The network-wide system becomes $\dot{\hat{x}}(t) = A\hat{x}(t)$ where $\hat{x}(t) = [\hat{x}_1^T(t), \cdots, \hat{x}_N^T(t)]^T$.

Since $\hat{x}_i^*(t)$ in (6.15) does not change the origin, the exponential stability of $\hat{x}_i(t)$ implies that the original system states $x_i(t)$, and inputs $u_i(t)$ in (6.1), (6.3), (6.4), (6.11) and (6.12) exponentially track their manifolds (6.5), (6.6) and $\vartheta_i(t)$.

6.3.3 Frequency stability guarantee

The following theorem summarizes the exponential stability of system states $x_i(t)$ under distributed internal model controller (6.11), (6.12) and (6.16) with respect to their manifolds (6.5), (6.6) and $\vartheta_i(t)$.

Theorem 6.3.1 *Consider distributed control law* (6.11)*,* (6.12) *and* (6.16)*. Under Assumptions 6.2.1 and 6.2.2, system states $x(t)$ are exponentially stable with respect to their manifolds* (6.5)*,* (6.6) *and $\vartheta_i(t)$ if matrix A is Hurwitz. In addition, there always exists a set of matrices M_i, N_i and gains $k_{i,1}, k_{i,2}, k_{i,3}$ such that matrix A is Hurwitz.*

In the proof, we provide Algorithm 8 to identify a set of gains and matrices in a distributed way such that A is Hurwitz.

6.4 Controller synthesis for robust adaptive frequency control

In this section, we study the case where the frequencies ρ_i in exosystem (6.2) are unknown. Internal models (6.11) and (6.12) will be used, but $T_i^{-1}(\rho_i)$ in (6.10) is uncertain to control authority i due to the unknown frequencies ρ_i. To address the challenge, we propose a new distributed adaptive internal model controller.

Let us define the augmented state $x_i(t) = [x_{i,1}(t),\ x_{i,2}(t), x_{i,3}^T(t)]^T = [w_i(t),\ P_{M_i}(t), \eta_i^T(t)]^T$ (or $x_i(t) = [x_{i,1}(t), x_{i,2}(t), x_{i,3}^T(t)]^T = [w_i(t), P_{C_i}(t), \eta_i^T(t)]^T$ for $i \in \mathcal{L}$) and corresponding manifolds $x_i^*(t)$.

6.4.1 Controller design

Like Section 6.3.2, a coordinate transformation is conducted to convert the global control problem into a global stabilization problem of the error dynamics. Also,

a backstepping approach is applied to ensure the stability of the error dynamics. Consider the transformation

$$\hat{x}_i(t) \triangleq x_i(t) - x_i^*(\Lambda_i(t)x_{i,2}(t), t) - \hat{x}_i^*(t) \tag{6.19}$$

where $\hat{x}_{i,1}^*(t) = 0, \hat{x}_{i,2}^*(t) = m_i N_i x_{i,1}(t)$. The first two terms in (6.19) define the tracking error and the third term is introduced by a backstepping technique to stabilize the error dynamics as (6.15). Consider inputs

$$\begin{aligned} P_{M_i}(t) &= P_{M_i}^*(\Lambda_i(t)x_{i,3}(t), t) - m_i(k_i - \frac{D_i}{m_i})x_{i,1}(t) \\ P_{C_i}(t) &= P_{C_i}^*(\Lambda_i(t)x_{i,3}(t), t) + m_i(k_i - \frac{D_i}{m_i})x_{i,1}(t). \end{aligned} \tag{6.20}$$

Under coordinate transformation (6.19), the frequency control problem is transformed to a stabilization problem for the same reason as (6.15). The difference is that we use estimated vector $\Lambda_i(t)$ instead of true vector $\Lambda_i^*(\rho_i) = \Psi_i T_i^{-1}(\rho_i)$. The origin of the error dynamics does not change by $\hat{x}_i^*(t)$.

Through coordinate transformation (6.19) and input (6.20), systems (6.4), (6.7) and internal model (6.11) become

$$\dot{\hat{x}}_i(t) = A_i(\Lambda_i^*(\rho_i))\hat{x}_i(t) + B_i(\hat{\Lambda}_i(t))x_{i,3}(t) \tag{6.21}$$

for $\forall i \in \mathcal{V}$ where $\hat{\Lambda}_i(t) \triangleq \Lambda_i(t) - \Lambda_i^*(\rho_i)$ is the estimation error and

$$\begin{aligned} A_i(\Lambda_i^*(\rho_i)) &= \begin{bmatrix} -k_i + \Lambda_i^*(\rho_i)N_i & \frac{1}{m_i}\Lambda_i^*(\rho_i) \\ (m_i M_i + D_i I)N_i & M_i \end{bmatrix}, \\ B_i(\hat{\Lambda}_i(t)) &= \begin{bmatrix} \frac{\hat{\Lambda}_i(t)}{m_i} \\ \mathbf{0}_{2\ell_i \times 2\ell_i} \end{bmatrix}. \end{aligned}$$

6.4.2 Projected parameter estimator

The quantity $\Lambda_i(t)$ is an estimate of $\Lambda_i^*(\rho_i)$ and its update law is given by:

$$\dot{\Lambda}_i^T(t) = J_i(t) - (\|J_i(t)\| + \gamma_i)(sgn(\Lambda_i(t) - \|\frac{(\rho_{\max}^2 + 1)\ell_i + \|M_i\|_F}{\|N_i\|}\|\mathbf{1}_{2\ell_i \times 1})/2$$

$$+ sgn(\Lambda_i(t) + \|\frac{(\rho_{\max}^2+1)\ell_i + \|M_i\|_F}{\|N_i\|}\|\mathbf{1}_{2\ell_i\times 1})/2) \tag{6.22}$$

where $J_i(t) = -\frac{\hat{x}_{i,1}(t)}{m_i}x_{i,3}(t)$ and $\gamma_i > 0$ is an arbitrary constant. The first term $J_i(t)$ in (6.22) is designed to cancel cross term $\hat{\Lambda}_i(t)\frac{\hat{x}_{i,1}(t)}{m_i}x_{i,3}(t)$ by $\hat{\Lambda}_i(t)\dot{\hat{\Lambda}}_i^T(t)$ in the Lyapunov analysis. The additional terms in (6.22) speed up the convergence rate by restricting the parameter estimates within $\|\Lambda_i(t)\| \leq \sqrt{2\ell_i}((\rho_{\max}^2+1)\ell_i + \|M_i\|_F)/\|N_i\|$. The bound is shown in Claim B in the proof of Theorem 6.4.1.

6.4.3 Frequency stability guarantee

The following theorem summarizes the asymptotic convergence of system states $x_i(t)$ to their manifolds (6.8) and $\vartheta_i(t)$. Consider matrix

$$\bar{A}_i = \begin{bmatrix} \bar{A}_i(1,1) & ((m_iM_i + D_iI)N_i)^T/2 \\ (m_iM_i + D_iI)N_i/2 & (M_i + M_i^T)/2 + 2I_{2\ell_i\times 2\ell_i} \end{bmatrix}$$
$$\bar{A}_i(1,1) = -k_i + (\rho_{\max}^2+1)\ell_i + \|M_i\|_F + ((\rho_{\max}^2+1)\ell_i + \|M_i\|_F)^2/(4m_i^2\|N_i\|^2). \tag{6.23}$$

Theorem 6.4.1 *Consider distributed control law* (6.11), (6.12) *and* (6.20) *and adaptive law* (6.22). *Under Assumptions 6.2.1 and 6.2.3, system states $x(t)$ are asymptotically convergent to their manifolds* (6.8) *and $\vartheta_i(t)$, if matrix $\bar{A}_i$ is negative definite for $\forall i \in \mathcal{V}$. In addition, there always exists a set of matrices M_i, N_i and gain k_i such that matrix $\bar{A}_i$ is negative definite.*

In the proof, we provide Algorithm 9 to identify a set of control gain and matrices in a distributed way such that $\bar{A}_i$ is negative definite.

Remark 6.4.1 *Simplified synchronous generator model* (6.7) *(as well as the simplified controllable load model) prevent potential problems where the adaptive law relies on unmeasurable values $\vartheta_i(t)$ and $T_i^{-1}(\rho_i)$ to eliminate cross terms.* ■

6.5 Analysis

This section presents the proofs of Theorem 6.3.1 and 6.4.1.

Algorithm 8 Distributed selection of control gains

for $i \in \mathcal{V}$ **do**
  Choose a controllable pair (M_i, C_i) such that M_i is Hurwitz and $\lambda_{\max}(\frac{M_i+M_i^T}{2}) < -3.5 - |\mathcal{N}_i|/(|\mathcal{N}_i|+2)^2$;
  Choose $0 < \alpha_i < \frac{2(-\lambda_{\max}(\frac{M_i+M_i^T}{2})-3.5-\frac{|\mathcal{N}_i|}{(|\mathcal{N}_i|+2)^2})^{\frac{1}{2}}}{\|(m_i M_i + D_i I_{2\ell_i})C_i\|}$;
  $N_i = \alpha_i C_i$;
  Find the solution $T_i^{-1}(\rho_i)$ of Sylvester equation (6.10);
end for
for $i \in \mathcal{V}$ **do**
  Choose $k_{i,1}, k_{i,2}, k_{i,3}$ sequentially such that
$k_{i,1} > \|\Psi_i T_i^{-1}(\rho_i)\|^2/(4m_i^2) + 1/(4m_i^2) + 1.5$,
$k_{i,2} > \frac{K_{m_i}^2}{4T_{CH_i}^2} + (e_i^* + k_{i,1})^2\|\Psi_i T_i^{-1}(\rho_i)\|^2/4 + 1.5$,
$k_{i,3} > T_{CH_i}^2(e_i^* + k_{i,1})^2(e_i^* + k_{i,2})^2\|\Psi_i T_i^{-1}(\rho_i)\|^2/(4K_{m_i}^2) + \sum_{j\in\mathcal{N}_i} t_{ij}^2(|\mathcal{N}_i| + 2)^2(\|\Psi_i T_i^{-1}(\rho_i)\|^2 + \|\Psi_j T_j^{-1}\|^2)/(4K_{m_i}^2) + 1.5$.
end for

6.5.1 Proof of Theorem 6.3.1

Assume that A is Hurwitz. Then, linear time invariance system $\dot{\hat{x}}(t) = A\hat{x}(t)$ is exponentially stable. Since coordinate transformation $\hat{x}_i^*(t)$ in (6.15) does not change the origin, this further implies that $x(t)$ in (6.3) is exponentially stable with respect to their manifolds (6.5), (6.6) and $\vartheta_i(t)$. One can prove the necessity part by reversing the steps above.

Now we proceed to prove the existence of control gains and matrices by construction. Consider system (6.17) where matrices and control gains are chosen by Algorithm 8. We will show that A is Hurwitz by verifying that the system is exponentially stable. Consider Lyapunov function candidate $V_i(t) = \frac{1}{2}\|\hat{x}_i(t)\|^2$ for $\forall i \in \mathcal{V}$. Since $\hat{x}_i^T(t)A_i\hat{x}_i(t) \in \mathbb{R}$, $\hat{x}_i^T(t)A_i\hat{x}_i(t) = (\hat{x}_i^T(t)A_i\hat{x}_i(t))^T$. Hence, the Lie derivative of Lyapunov function candidate along the trajectories of system (6.17) becomes

$$\begin{aligned}\dot{V}_i(t) &= \hat{x}_i^T(t)A_i\hat{x}_i(t) + \sum_{j\in\mathcal{N}_i}\hat{x}_i(t)B_{ij}\hat{x}_{j,4}(t)\\ &= \hat{x}_i^T(t)\frac{A_i + A_i^T}{2}\hat{x}_i(t) + \sum_{j\in\mathcal{N}_i}\hat{x}_i(t)B_{ij}\hat{x}_{j,4}(t).\end{aligned}$$

Since $\hat{x}_i(t)B_{ij}\hat{x}_{j,4}(t) \leq \frac{\delta}{2}\|\hat{x}_{i,3}(t)\|^2 + \frac{\|B_{ij}\|^2}{2\delta}\|\hat{x}_{j,4}(t)\|^2$ with $\delta = \frac{t_{ij}}{K_{m_i}}(|\mathcal{N}_i|+2)^2\|\Psi_j T_j^{-1}(\rho_j)\|/2$, we have

$$\dot{V}_i(t) \leq \hat{x}_i^T(t)\bar{A}_i\hat{x}_i(t) + \sum_{j\in\mathcal{N}_i}\hat{x}_j^T(t)\bar{B}_{ij}\hat{x}_j(t)$$

where

$$\bar{A}_i = \frac{A_i + A_i^T}{2} + \begin{bmatrix} \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times 1} & \mathbf{0}_{2\times 2\ell_i} \\ \mathbf{0}_{1\times 2} & P_i(2,2) & \mathbf{0}_{1\times 2\ell_i} \\ \mathbf{0}_{2\ell_i\times 2} & \mathbf{0}_{2\ell_i\times 1} & \mathbf{0}_{2\ell_i\times 2\ell_i} \end{bmatrix}$$

and $P_i(2,2) = \sum_{j\in\mathcal{N}_i}\frac{t_{ij}^2}{K_{m_i}^2}(|\mathcal{N}_i|+2)^2\|\Psi_j T_j^{-1}(\rho_j)\|^2/4$ for $i \in \mathcal{G}$ and $\bar{A}_i = \frac{A_i + A_i^T}{2}$ for $i \in \mathcal{L}$, and $\|\bar{B}_{ij}\| \leq \frac{1}{(|\mathcal{N}_i|+2)^2}$.

Claim A: It holds that $\hat{x}_i^T(t)\bar{A}_i\hat{x}_i(t) \leq -0.5\|\hat{x}_i(t)\|^2$.

PROOF. Since $\hat{x}_{i,l}^T(t)\bar{A}_i(l,p)\hat{x}_{i,p}(t) \leq \frac{\delta}{2}\|\hat{x}_{i,l}(t)\|^2 + \frac{\|\bar{A}_i(l,p)\|^2}{2\delta}\|\hat{x}_{i,p}(t)\|^2$ for any $\delta > 0$ and $\hat{x}_{i,4}^T(t)M_i\hat{x}_{i,4}(t) = \hat{x}_{i,4}^T(t)(\frac{M_i+M_i^T}{2})\hat{x}_{i,4}(t) \leq \lambda_{\max}(\frac{M_i+M_i^T}{2})\|\hat{x}_{i,4}(t)\|^2$ by the Rayleigh quotient [162], we have

$$\hat{x}_i^T(t)\bar{A}_i\hat{x}_i(t) \leq \hat{x}_i^T(t)A_i'\hat{x}_i(t) \tag{6.24}$$

where $A_i' = \text{diag}(A_i'(1,1), A_i'(2,2), A_i'(3,3), A_i'(4,4))$ for $i \in \mathcal{G}$ and $A_i' = \text{diag}(A_i'(1,1), A_i'(2,2), A_i'(4,4))$ for $i \in \mathcal{L}$, and

$$\begin{aligned}
A_i'(1,1) &= -k_{1,k} + \frac{\|\Psi_i T_i^{-1}(\rho_i)\|^2}{4m_i^2} + \frac{1}{4m_i^2} + 1 \\
A_i'(2,2) &= -k_{2,k} + \frac{K_{m_i}^2}{4T_{CH_i}^2} + \frac{(e_i^* + k_{i,1})^2\|\Psi_i T_i^{-1}(\rho_i)\|^2}{4} + 1 \\
A_i'(3,3) &= -k_{3,k} + 1 + T_{CH_i}^2(e_i^* + k_{i,1})^2(e_i^* + k_{i,2})^2\|\Psi_i T_i^{-1}(\rho_i)\|^2/(4K_{m_i}^2) \\
&\quad + \sum_{j\in\mathcal{N}_i}\frac{t_{ij}^2}{4K_{m_i}^2}(|\mathcal{N}_i|+2)^2(\|\Psi_i T_i^{-1}(\rho_i)\|^2 + \|\Psi_j T_j^{-1}(\rho_j)\|^2) \\
A_i'(4,4) &= (\lambda_{\max}(\frac{M_i + M_i^T}{2}) + \frac{\|(m_i M_i + D_i I_{2\ell_i})N_i\|^2}{4} + 3 + |\mathcal{N}_i|/(|\mathcal{N}_i|+2)^2)I_{2\ell_i}.
\end{aligned}$$

Algorithm 8 ensures $A_i'(l,l) < -0.5$, and thus $\lambda_{\max}(A_i') < -0.5$ because A_i' is a

diagonal matrix. Hence, by (6.24) and the Rayleigh quotient,

$$\hat{x}_i^T(t)\bar{A}_i\hat{x}_i(t) \leq \lambda_{\max}(A_i')\|\hat{x}_i(t)\|^2 \leq -0.5\|\hat{x}_i(t)\|^2.$$

■

By Claim A,

$$\dot{V}_i(t) \leq -V_i(t) + \sum_{j\in\mathcal{N}_i} \frac{\|\hat{x}_j(t)\|^2}{(|\mathcal{N}_i|+2)^2}.$$

Consider $U_i(t) = \frac{1}{2}\|\hat{x}_i(t)\|^2$ and $\dot{U}_i(t) = -U_i(t) + \sum_{j\in\mathcal{N}_i} \frac{\|\hat{x}_j(t)\|^2}{(|\mathcal{N}_i|+2)^2}$. By the comparison lemma (Lemma 3.4 [163]), it holds that $V_i(t) \leq U_i(t)$ for $t \geq 0$ when $V_i(0) \leq U_i(0)$. The general solution $U_i(t)$ of the linear differential equation satisfies

$$\begin{aligned}
U_i(t) &= e^{-t}U_i(0) + \sum_{j\in\mathcal{N}_i} \int_0^t e^{-(t-\tau)} \frac{\|\hat{x}_j(\tau)\|^2}{(|\mathcal{N}_i|+2)^2} d\tau \\
&\leq e^{-t}U_i(0) + \sum_{j\in\mathcal{N}_i} \frac{\|\hat{x}_j\|^2_{[0,t]}}{(|\mathcal{N}_i|+2)^2} \int_0^t e^{-(t-\tau)} d\tau \\
&\leq e^{-t}U_i(0) + \sum_{j\in\mathcal{N}_i} \frac{\|\hat{x}_j\|^2_{[0,t]}}{(|\mathcal{N}_i|+2)^2} (1 - e^{-t}) \\
&\leq e^{-t}U_i(0) + \sum_{j\in\mathcal{N}_i} \frac{\|\hat{x}_j\|^2_{[0,t]}}{(|\mathcal{N}_i|+2)^2}.
\end{aligned}$$

Given $V_i(0) = U_i(0)$, it follows from $V_i(t) \leq U_i(t)$ that

$$V_i(t) \leq e^{-t}V_i(0) + \sum_{j\in\mathcal{N}_i} \frac{\|\hat{x}_j\|^2_{[0,t]}}{(|\mathcal{N}_i|+2)^2}.$$

By taking norm, Cauchy-schwarz inequality and square-root to the above inequality, we have

$$\begin{aligned}
\|\hat{x}_i(t)\| &\leq e^{-0.5t}\|\hat{x}(0)\| + \sum_{j\in\mathcal{N}_i} \frac{\|\hat{x}_j\|_{[0,t]}}{|\mathcal{N}_i|+2} \\
&\leq \max\{(|\mathcal{N}_i|+1)e^{-0.5t}\|\hat{x}_i(0)\|, \frac{|\mathcal{N}_i|+1}{|\mathcal{N}_i|+2} \max_{j\in\mathcal{N}_i}\{\|\hat{x}_j\|_{[0,t]}\}\}. \qquad (6.25)
\end{aligned}$$

Inequality (6.25) implies that $\hat{x}_i(t)$ is input-to-state stable (ISS) [164] with respect

to each $\hat{x}_j(t)$ with a contractive linear gain. By distributed constrained small-gain theorem 6.8.1, $\hat{x}(t)$ is exponentially stable. The exponential stability of $\hat{x}(t)$ implies that matrix A is Hurwitz. ■

6.5.2 Proof of Theorem 6.4.1

Consider Lyapunov function candidate

$$V(t) = \frac{1}{2}\|\hat{x}(t)\|^2 + \frac{1}{2}\sum_{i\in\mathcal{V}}\|\hat{\Lambda}_i(t)\|^2.$$

The Lie derivative of Lyapunov function candidate along the trajectories of system (6.21) becomes

$$\dot{V}(t) = \hat{x}^T(t)A(\Lambda^*(\rho))\hat{x}(t) + \sum_{i\in\mathcal{V}}\hat{\Lambda}_i(t)(\dot{\hat{\Lambda}}_i^T(t) - J_i(t)) \tag{6.26}$$

where $A(\Lambda^*(\rho)) = \text{diag}(A_1(\Lambda_i^*(\rho_i)), \cdots, A_{|\mathcal{V}|}(\Lambda_i^*(\rho_i)))$. Claim B identifies an upper bound of uncertain term $\Lambda_i^*(\rho_i)$.

Claim B: $\|\Lambda_i^*(\rho_i)\| \leq ((\rho_{\max}^2 + 1)\ell_i + \|M_i\|_F)/\|N_i\|$.

PROOF. By post-multiplying $T_i^{-1}(\rho_i)$ and taking Frobenius norm on both sides of Sylvester equation (6.10), we have

$$\|N_i\Lambda_i^*(\rho_i)\|_F \leq \|T_i(\rho_i)\Phi_i(\rho_i)T_i^{-1}(\rho_i)\|_F + \|M_i\|_F. \tag{6.27}$$

By the definition of Frobenius norm, it holds that

$$\begin{aligned}\|N_i\Lambda_i^*(\rho_i)\|_F &= \sqrt{\sum_{p=1}^{2\ell_i}\sum_{l=1}^{2\ell_i} N_{i,p}^2(\Lambda_{i,l}^*(\rho_i))^2} \\ &= \sqrt{(\sum_{p=1}^{2\ell_i} N_{i,p}^2)(\sum_{l=1}^{2\ell_i}(\Lambda_{i,l}^*(\rho_i))^2)} = \|N_i\|_F\|\Lambda_i^*(\rho_i)\|_F.\end{aligned} \tag{6.28}$$

By (6.28) and $\|\cdot\|_F \leq \|\cdot\|_{tr}$ (Lemma 10 in [165]), (6.27) becomes

$$\begin{aligned}\|N_i\|_F\|\Lambda_i^*(\rho_i)\|_F &\leq \|T_i(\rho_i)\Phi_i(\rho_i)T_i^{-1}(\rho_i)\|_{tr} + \|M_i\|_F \\ &= \|\Phi_i(\rho_i)\|_{tr} + \|M_i\|_F \leq (\rho_{\max}^2 + 1)\ell_i + \|M_i\|_F.\end{aligned}$$

Note that $\|\cdot\|_F = \|\cdot\|_2$ for a vector. ■

Claim C shows that $\hat{\Lambda}_i(t)(\dot{\hat{\Lambda}}_i^T(t) - J_i(t))$ is non-positive.

Claim C: $\hat{\Lambda}_i(t)(\dot{\hat{\Lambda}}_i^T(t) - J_i(t)) \leq 0$.

PROOF. If $|\hat{\Lambda}_{i,l}(t)| < \|\frac{(\rho_{\max}^2+1)\ell_i+\|M_i\|_F}{\|N_i\|}\|$ for all l, then $\hat{\Lambda}_i(t)(\dot{\hat{\Lambda}}_i^T(t) - J_i(t)) = 0$. Assume there exists l such that $|\hat{\Lambda}_{i,l}(t)| \geq \|\frac{(\rho_{\max}^2+1)\ell_i+\|M_i\|_F}{\|N_i\|}\|$. Let $S_{i,l}^+(t)$ denote a set of indices l such that $\hat{\Lambda}_{i,l}(t) \geq \|\frac{(\rho_{\max}^2+1)\ell_i+\|M_i\|_F}{\|N_i\|}\|$. Likewise, $S_{i,l}^-(t) \triangleq \{l|\hat{\Lambda}_{i,l}(t) \leq -\|\frac{(\rho_{\max}^2+1)\ell_i+\|M_i\|_F}{\|N_i\|}\|\}$. Then, we have

$$\begin{aligned}\hat{\Lambda}_i(t)(\dot{\hat{\Lambda}}_i^T(t) - J_i(t)) &\leq (\sum_{l\in S_{i,l}^+(t)} -\hat{\Lambda}_{i,l}(t)(\|J_i(t)\| + \gamma_i) + \sum_{l\in S_{i,l}^-(t)} \hat{\Lambda}_{i,l}(t)(\|J_i(t)\| + \gamma_i)) \\ &\leq -\gamma_i(\sum_{l\in S_{i,l}^+(t)} |\hat{\Lambda}_{i,l}(t)| + \sum_{l\in S_{i,l}^-(t)} |\hat{\Lambda}_{i,l}(t)|) \leq 0\end{aligned}$$

where $\hat{\Lambda}_{i,l}(t) = |\hat{\Lambda}_{i,l}(t)|$ for $l \in S_{i,l}^+(t)$ and $\hat{\Lambda}_{i,l}(t) = -|\hat{\Lambda}_{i,l}(t)|$ for $l \in S_{i,l}^-(t)$ are applied. ■

By Claim B, we have

$$\begin{aligned}\|\hat{x}_{i,1}^T(t)\Lambda_i^*(\rho_i)\hat{x}_{i,3}(t)\| &\leq \|\hat{x}_{i,1}^T(t)\|\|\Lambda_i^*(\rho_i)\hat{x}_{i,3}(t)\| \\ &\leq \frac{(\rho_{\max}^2+1)\ell_i + \|M_i\|_F}{\|N_i\|}(\frac{\delta\|\hat{x}_{i,1}(t)\|^2}{2} + \frac{\|\hat{x}_{i,3}(t)\|^2}{2\delta})\end{aligned}$$

and then

$$\hat{x}^T(t)A(\Lambda^*(\rho))\hat{x}(t) \leq \hat{x}^T(t)\bar{A}\hat{x}(t) \tag{6.29}$$

Symmetric matrix $\bar{A} = \text{diag}(\bar{A}_1, \cdots, \bar{A}_{|\mathcal{V}|})$ is negative definite where $\bar{A}_i$ is defined in (6.23). Thus, Claim C and (6.29) lead (6.26) to

$$\dot{V}(t) \leq \hat{x}^T(t)\bar{A}\hat{x}(t) \leq \lambda_{\max}(\bar{A})\|\hat{x}(t)\|^2. \tag{6.30}$$

Take the integral from 0 to t on both sides of (6.30), then

$$-\lambda_{\max}(\bar{A})\int_0^t \|\hat{x}(\tau)\|^2 d\tau \leq -\int_0^t \dot{V}(\tau)d\tau \leq V(0) < \infty$$

where $V(t) \geq 0$ is applied. Since $\int_0^t \|\hat{x}(\tau)\|^2 d\tau$ is non-decreasing and upper bounded by $-V(0)/\lambda_{\max}(\bar{A})$, the limit $\lim_{t\to\infty}\int_0^t \|\hat{x}(\tau)\|^2 d\tau$ exists. Moreover, $\|\hat{x}(t)\|^2$ is

uniformly continuous as shown in Claim D.

Claim D: $\|\hat{x}(t)\|^2$ is uniformly continuous.

PROOF. Recall $\dot{V}(t)$ is non-positive. There exists a constant $U > 0$ such that $\|\hat{x}(t)\| \leq U$ for $t \in [0, \infty)$. Consider

$$\begin{aligned} |\|\hat{x}(t+s)\|^2 - \|\hat{x}(t)\|^2| &= \sum_{i\in\mathcal{V}} |\sum_{l=1}^{3} (\|\hat{x}_{i,l}(t+s)\|^2 - \|\hat{x}_{i,l}(t)\|^2)| \\ &\leq \sum_{i\in\mathcal{V}} |\sum_{l=1}^{3} |\|\hat{x}_{i,l}(t+s)\|^2 - \|\hat{x}_{i,l}(t)\|^2| \ |. \end{aligned} \tag{6.31}$$

The term $\hat{x}_{i,l}(t+s)$ is given by $\hat{x}_{i,l}(t+s) = \hat{x}_{i,l}(t) + \int_t^{t+s} \dot{\hat{x}}_{i,l}(\tau) d\tau$. By uniform boundedness of all $\hat{x}_{i,l}(t)$, for $l = 1, 2$ and any $s > 0$, we have

$$\hat{x}_{i,l}(t) - a_{i,l} s \leq \hat{x}_{i,l}(t+s) \leq \hat{x}_{i,l}(t) + a_{i,l} s$$

where $a_{i,1}$ is a positive constant and $a_{i,2} = a[1, \cdots, 1]^T$ is a vector with a positive constant a. Therefore,

$$\begin{aligned} |\|\hat{x}_{i,l}(t+s)\|^2 - \|\hat{x}_{i,l}(t)\|^2| &\leq \|2a_{i,l}^T \hat{x}_{i,l}(t) s\| + \|a_{i,l}^T a_{i,l} s^2\| \\ &\leq \|2a_{i,l} s\| U + \|a_{i,l}^T a_{i,l} s^2\| \end{aligned}$$

where the right hand side is strictly increasing in s and $\lim_{s\to 0} \|2a_{i,l}s\|U + \|a_{i,l}^T a_{i,l} s^2\| = 0$. By applying the above bound to (6.31), we can prove the uniform continuity of $\|x(t)\|^2$; i.e., for any $\epsilon > 0$, there always exists $\delta > 0$ such that for all t and $0 \leq s \leq \delta$, $|\|x(t+s)\|^2 - \|x(t)\|^2| \leq \epsilon$. ■

It has been shown that $\|\hat{x}(t)\|^2$ is uniformly continuous, and $\lim_{t\to\infty} \int_0^t \|\hat{x}(\tau)\|^2 d\tau$ exists and is finite. By the Barbalat's lemma (Lemma 8.2 in [163]), $\|\hat{x}(t)\|^2$ asymptotically converges to zero.

Now we proceed to prove the existence of matrices and control gain such that matrix $\bar{A}_i$ is negative definite by construction. Consider a set of matrices and control gain by Algorithm 9. Since $\hat{x}_{i,l}^T(t) \bar{A}_i(l,p) \hat{x}_{i,p}(t) \leq \frac{\delta}{2} \|\hat{x}_{i,l}(t)\|^2 + \frac{\|\bar{A}_i(l,p)\|^2}{2\delta} \|\hat{x}_{i,p}(t)\|^2$ for any $\delta > 0$ and $\hat{x}_{i,2}^T(t) M_i \hat{x}_{i,2}(t) = \hat{x}_{i,2}^T(t) \frac{M_i + M_i^T}{2} \hat{x}_{i,2}(t) \leq \lambda_{\max}(\frac{M_i + M_i^T}{2}) \|\hat{x}_{i,2}(t)\|^2$, we have

$$\hat{x}_i^T(t) \bar{A}_i \hat{x}_i(t) \leq \hat{x}_i^T(t) A_i' \hat{x}_i(t)$$

Algorithm 9 Distributed selection of control gains

1: **for** $i \in \mathcal{V}$ **do**
2: Choose a controllable pair (M_i, C_i) such that M_i is Hurwitz and $\lambda_{\max}(\frac{M_i+M_i^T}{2}) < -1$;
3: Choose $0 < \alpha_i < \frac{2(-\lambda_{\max}(\frac{M_i+M_i^T}{2})-1)^{\frac{1}{2}}}{\|(m_iM_i+D_iI_{2\ell_i})C_i\|}$;
4: $N_i = \alpha_i C_i$;
5: Choose k_i such that $k_i > (\rho_{\max}^2+1)\ell_i + \frac{((\rho_{\max}^2+1)\ell_i+\|M_i\|_F)^2}{4m_i^2\|N_i\|^2} + 1 + \|M_i\|_F$.
6: **end for**

where $A_i' = \mathrm{diag}(A_i'(1,1), A_i'(2,2))$ and $A_i'(1,1) = -k_i + (\rho_{\max}^2+1)\ell_i + \|M_i\|_F + 1 + \frac{((\rho_{\max}^2+1)\ell_i+\|M_i\|_F)^2}{4m_i^2\|N_i\|^2}$, $A_i'(2,2) = (\lambda_{\max}(\frac{M_i+M_i^T}{2}) + \frac{\|(m_iM_i+ID_i)N_i\|^2}{4} + 1)I_{2\ell_i}$. Algorithm 9 ensures $A_i'(l,l) < 0$, and thus diagonal matrix A_i' is negative definite. This implies that $\bar{A}_i$ is also negative definite. ■

6.6 Simulation

In this section, we present simulations to show the performance of the proposed distributed controllers. Although the proposed controllers are designed under the assumption that voltages remain constants, we have taken into account voltage dynamics through the following power flow model (simplified from (6.101) in [105]):

$$\begin{aligned} P_{ij}(t) &= |V_i(t)||V_j(t)|B_{ij}\sin(\theta_i(t)-\theta_j(t)) \\ Q_{ij}(t) &= |V_i(t)||V_j(t)|(-B_{ij}\cos(\theta_i(t)-\theta_j(t))). \end{aligned} \tag{6.32}$$

Variables $P_{ij}(t)$ and $Q_{ij}(t)$ are active power flow and reactive power flow, respectively. Coefficient B_{ij} is an element in Y-matrix.

There are five simulations including 1. No fault case, 2. Three-phase fault, 3. Load-switching (light load), 4. Load-switching (peak load) and 5. Minni-WECC system [166–168]. The first four simulations are applied to the single line diagram of the IEEE 68-bus test system topology shown in [169, 170], and the last simulation is applied to the Minni-WECC system. All of the parameters are adopted from [85, 105]. We assume that each generator/load bus $i \in \mathcal{G}, \mathcal{L}$ has (unknown) local net load $P_{L_i}(t) = 0.05\sin(0.1t) + 0.05\sin(0.2t)$ with $\Psi_i = [1,0,1,0]$. Frequency upper bound is given by $\rho_{\max} = 0.9$.

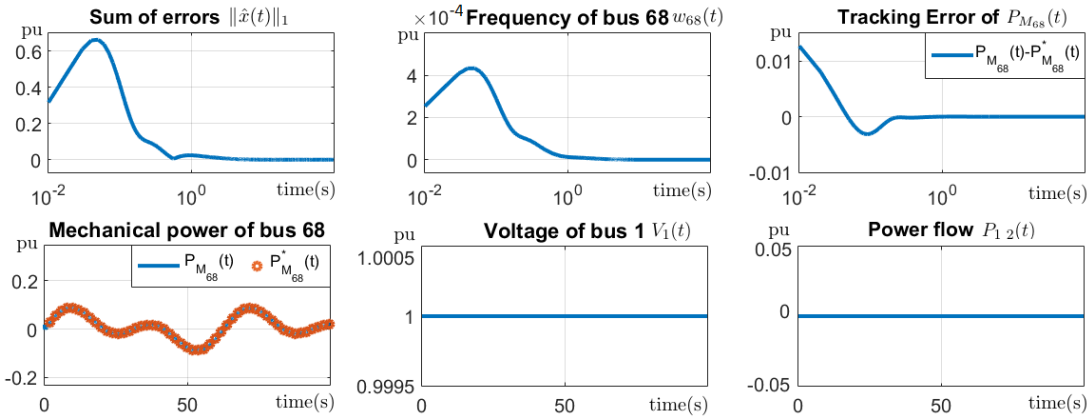

Figure 6.1: Case1: Simulation results of no fault case.

System and controller parameters. The generator parameters are adopted from page 598 in [105]: $R_i = 0.05$, $T_{CH_i} = 0.3$, $T_{G_i} = 0.2$, $K_{m_i} = 1$ and $K_{e_i} = 1$ for $\forall i \in \mathcal{G}$, and $m_i = 10$, $D_i = 1$ and $B_{ij} = 1.5$ for $\forall i \in \mathcal{V}$. Demand response parameters $b_i = (40\$/MWh)/(150s)$ and $c_i = (-0.8\$/MW^2h)/(150s)$ for $i \in \mathcal{L}$ are borrowed from [85].

For the *robust frequency control*, we choose $k_i = [1, 26, 99]^T$ and matrices

$$M_i = \begin{bmatrix} -5.9 & -2.1 & -0.1 & 1.5 \\ 2.4 & -6.3 & -0.2 & 2.9 \\ 0.8 & 0.9 & -6.6 & 2.5 \\ 1.6 & 0.3 & 0.8 & -7 \end{bmatrix}, \; N_i = \begin{bmatrix} 0.11 \\ -0.1 \\ 0.12 \\ 0.12 \end{bmatrix}$$

which satisfy that matrix A is Hurwitz in Theorem 6.3.1. For the *robust adaptive frequency control* problem, we choose $k_i = 45.5$ and the above matrices, which guarantee that $\bar{A}_i$ is negative definite in Theorem 6.4.1.

1. No fault. The first simulation is designed to show error tracking performance of the proposed controllers. The initial condition of the current simulation is intentionally chosen larger than that of the other simulations. Figure 6.1 summarizes the results for the *robust frequency control*. In each subfigure, the horizontal axis represents time in log-scale or linear-scale, and the vertical axis represents corresponding values in per unit. The first subfigure shows that the total state errors $\|\hat{x}(t)\|_1$ are exponentially stable; i.e., the designed distributed controller achieves the objective and eventually steers network-wide frequency deviations to 0. This

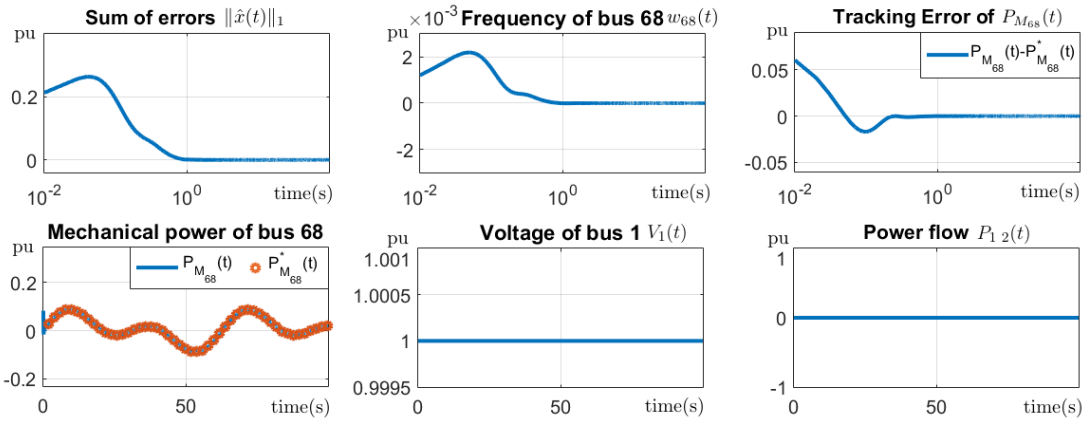

Figure 6.2: Case2: Simulation results of no fault case.

implies that the trajectories of controllable load $P_{C_i}(t)$ of bus $i \in \mathcal{L}$ and mechanical power $P_{M_i}(t)$ for $i \in \mathcal{G}$ track desired signals; e.g., mechanical power $P_{M_{68}}(t)$ for $68 \in \mathcal{G}$ tracks net loads as shown in the third and fourth subfigures. Moreover, all the frequency errors $\Delta w_i(t)$ are stable; e.g., frequency of bus 68 is shown in the second subfigure.

Figure 6.2 summarizes the results for the *robust adaptive frequency control*. The first subfigure shows that network-wide state errors converge to zero, implying that the angular frequencies are controlled to $60Hz$; e.g., the second subfigure, and $P_{C_i}(t)$ and $P_{M_i}(t)$ balance the local power demand and generation (as well as incoming and outgoing power).

The transient performance of the both cases in Figure 6.1 and 6.2 look similar to each other, while their theoretic guarantees in the theorems are different. The theoretic guarantees are valid for the worst case. That is, no matter what system parameters are, the robust controller always ensures exponential stability and the robust adaptive controller always ensures asymptotic stability. However, there could be some instances where the robust adaptive controller performs as good as or even better than the robust controller. The simulation in the paper is actually one of these cases and two controllers both achieve exponential stability. These cases do not violate the theorems.

2. Three-phase fault. In this simulation, bus $1 \in \mathcal{L}$ is assumed to have a three-phase fault with normal-clearing time (5 cycles). To simulate the three-phase fault, we set the corresponding voltage of bus 1 to a very low value; i.e.,

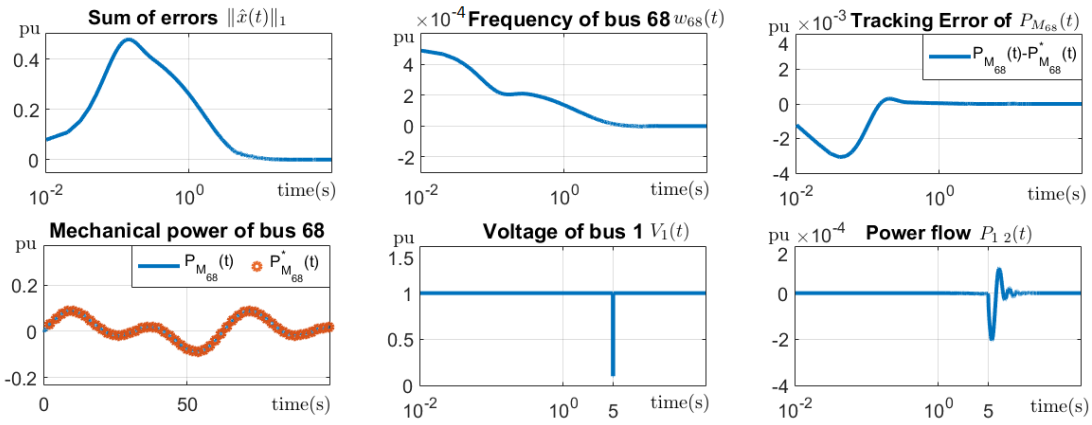

Figure 6.3: Case1: Three-phase fault at $t = 5$.

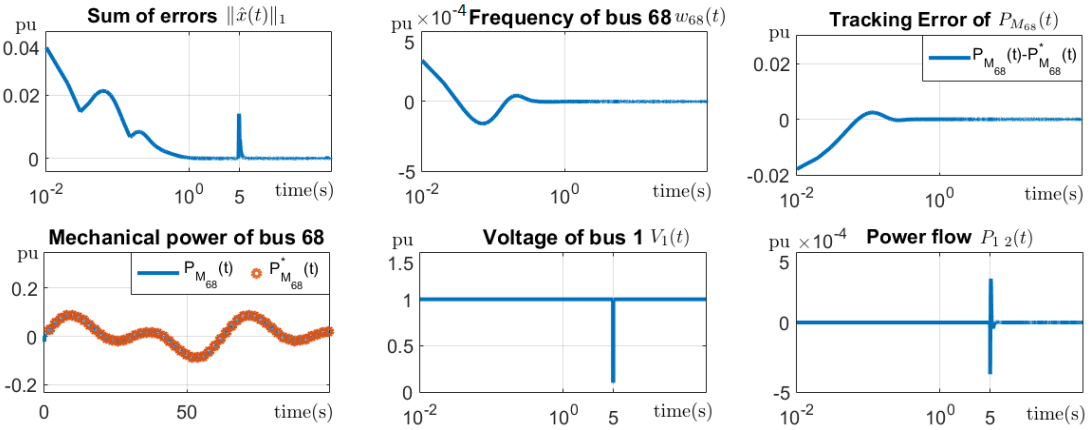

Figure 6.4: Case2: Three-phase fault at $t = 5$.

$V_1(t) = 0.1$ p.u. for $t \in [5, 5 + \frac{5}{60}]$. We present the results in Figures 6.3 and 6.4.

The fault induces a sudden increase in errors at $t = 5$ but the system restores stability without large deviations. This is because all local controllers are designed to collaborate together to stabilize the system by reducing their local errors.

3. Load-switching (light load). In this simulation, bus 1 suddenly disconnects a part of the load (0.5 p.u.) as well as net loads $P_{L_1}(t)$ for $t > 5$. At the same time, bus 1 loses control of the price. Figures 6.5 and 6.6 represent the results. Load-switching and loss of pricing control induce persistent errors in voltages as shown in the firth subfigure of Figures 6.5 and 6.6. This then increases total errors in the first subfigure. Voltages are slowly restored in Case 1, while they remain constant in Case 2. Although the local controllers are not particularly designed to

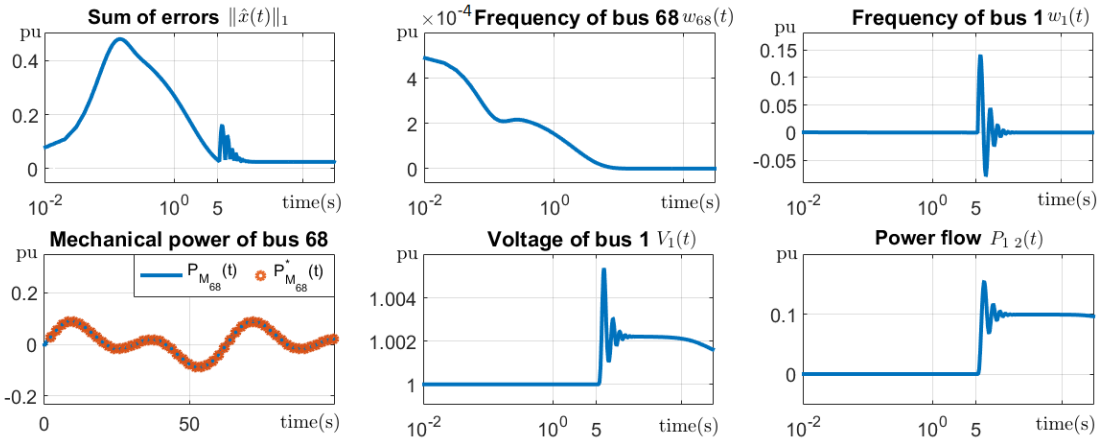

Figure 6.5: Case1: Load-switching at $t = 5$ during light load.

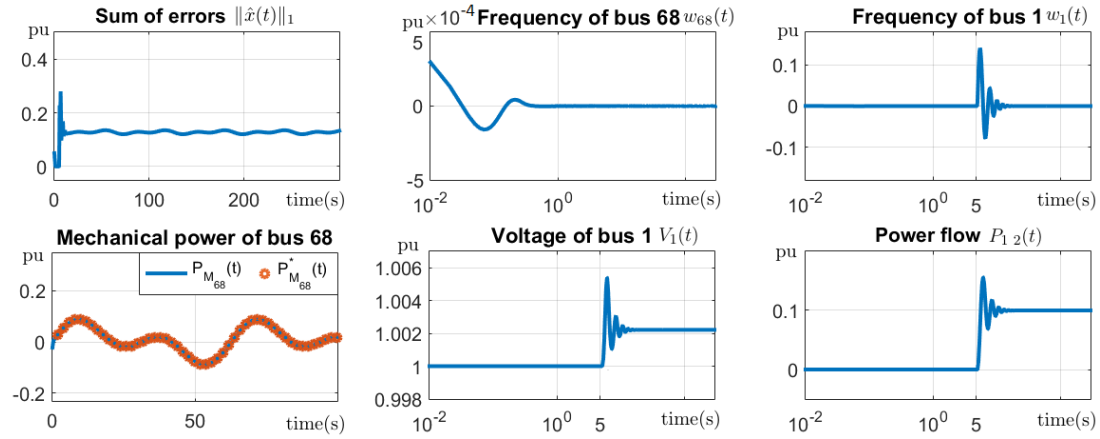

Figure 6.6: Case2: Load-switching at $t = 5$ during light load.

reduce errors of the neighboring buses, controllers in the neighboring buses of bus 1 collaborate to reduce errors of bus 1 through adaptively changing tie-line flow $P_{1j}(t)$.

4. Load-switching (peak load). We simulate the same load-switching case as simulation 3; i.e., bus 1 loses net load (0.5 p.u.) and loses pricing control. On top of this, bus 1 continuously has net loads $P_{L_i}(t)$, and, in the other buses, the amount of controllable load $P_{C_i}(t)$ and mechanical power $P_{M_i}(t)$ are restricted to $[-0.12, 0.12]$ p.u. Figures 6.7 and 6.8 show the results of this simulation. The operating condition is worse than that of simulation 3, but the distributed controllers successfully restrict the errors in low level as shown in the first subfigure of Figures 6.7 and 6.8. Overall oscillations come from persistent net loads in bus

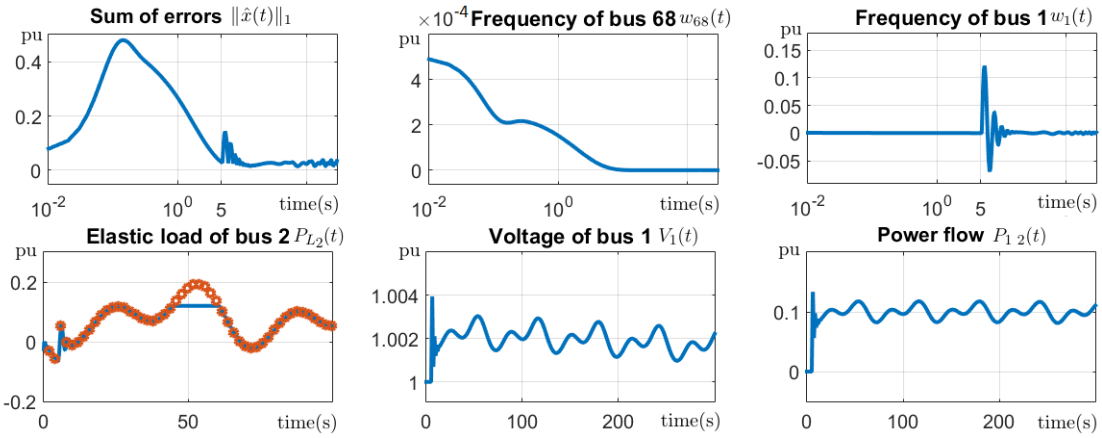

Figure 6.7: Case1: Load-switching at $t = 5$ during peak load.

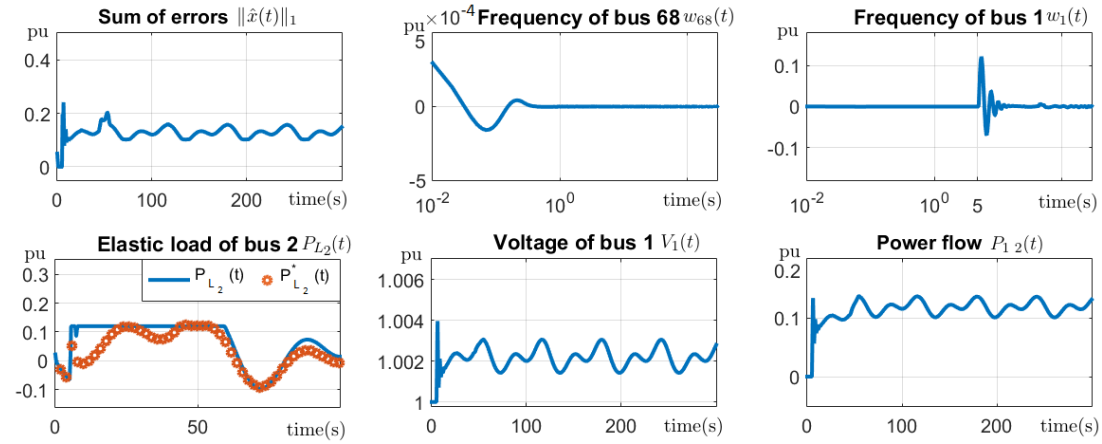

Figure 6.8: Case2: Load-switching at $t = 5$ during peak load.

1, which are not rejected and act as a source of unknown disturbances.

5. Minni-WECC system. The Minni-WECC model [166–168] is a reduced-order model of western electricity coordinating council (WECC) by applying generator equivalences and merging transmission paths. We conduct a no fault simulation on the Minni-WECC model, and the results are shown in Figures 6.9 and 6.10.

The results are similar to that of simulation 1. The proposed distributed controllers do not distinguish the topologies of the systems. Thus, the performance of the controllers remain similar even when the network changes.

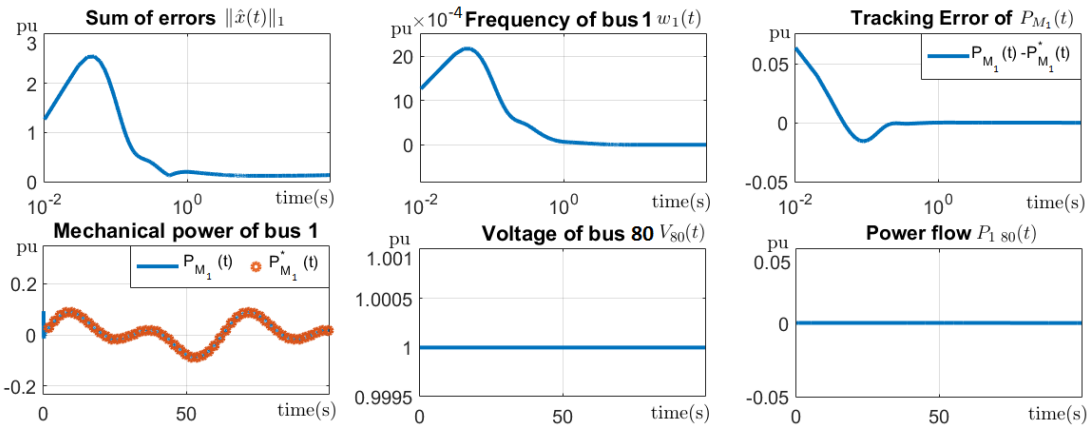

Figure 6.9: Case1: Simulation on the Minni-WECC model.

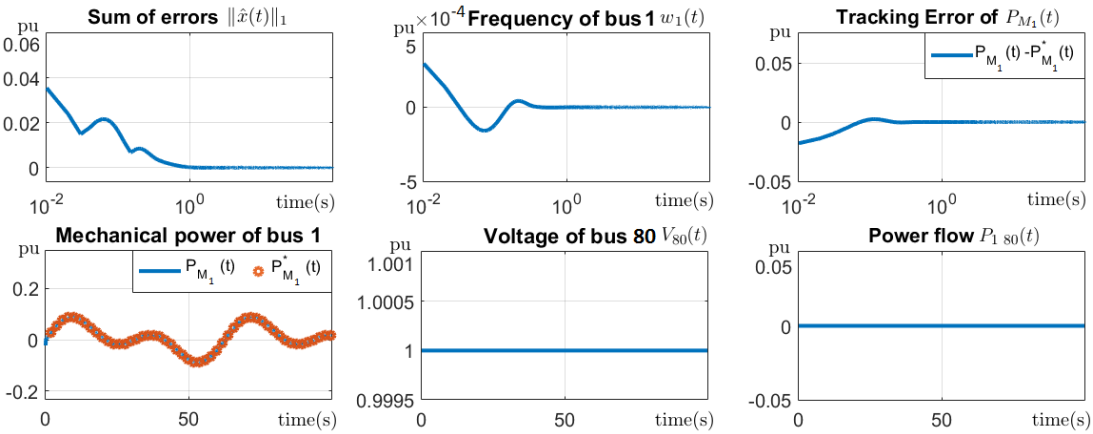

Figure 6.10: Case2: Simulation on the Minni-WECC model.

6.7 Conclusion

We have investigated the frequency control of multi-machine power systems subject to uncertain and dynamic net loads. The proposed distributed internal model controllers coordinate synchronous generators and demand response to ensure frequency stability. Simulations on the IEEE 68-bus test system demonstrate the performance of the controllers.

6.8 Appendix: Distributed constrained small-gain theorem

Distributed constrained small-gain theorem is introduced in this appendix. The theorem is an extension of constrained small-gain theorem in [171] to a network set-up.

Consider an undirected graph $(\mathcal{V}, \mathcal{E})$ and set $\mathcal{N}_i \triangleq \{j \in \mathcal{V} \setminus \{i\} | (i,j) \in \mathcal{E}\}$. The dynamic system associated with node i is given by

$$\dot{x}_i(t) = f_i(x(t), d_i(t), t) \tag{6.33}$$

where $x_i(t)$ and $d_i(t)$ denote system state and uncertainty respectively.

Assumption 6.8.1 *The system* (6.33) *is input-to-state stable with respect to neighboring states. Equivalently, there exist class $\mathcal{KL}$ function β_i and class $\mathcal{K}$ functions γ_{id} and γ_{ij} such that for $\forall t \geq t_0$ and $\forall i \in \mathcal{V}$,*

$$\|x_i(t)\| \leq \max\{\beta_i(\|x_i(t_0)\|, t - t_0), \gamma_{id}(\|d_i\|_{[t_0,t]}), \max_{j \in \mathcal{N}_i}\{\gamma_{ij}(\|x_j\|_{[t_0,t]})\}\}. \tag{6.34}$$

Assumption 6.8.2 *Gain functions γ_{ij} are contraction mappings for $(i,j) \in \mathcal{E}$; i.e., $\gamma_{ij}(s) < s$ for all $s > 0$.*

Theorem 6.8.1 ***(Distributed constrained small-gain theorem)*** *Under Assumptions 6.8.1 and 6.8.2, the system* (6.33) *is ISS with respect to d. Equivalently, there exists class $\mathcal{KL}$ function β and class $\mathcal{K}$ function γ_{id} such that for all $x_i(t_0) \in \hat{X}_i$ and $\|d\|_{[t_0,\infty)} < \hat{\Delta}_d$, the solution of* (6.33) *exists and for $\forall t \geq t_0$,*

$$\|x(t)\| \leq \max\{\beta(\|x(t_0)\|, t - t_0), \gamma_{id}(\|d(t)\|_{[t_0,t]}). \tag{6.35}$$

Moreover, the function $\beta(x,t) = |\mathcal{V}| \sum_{i \in \mathcal{V}} \beta_i(|\mathcal{V}| \sum_{k \in \mathcal{V}} \beta_k(x, 0), \frac{t}{(2\mathcal{L})^{|\mathcal{V}|-1}})$ is a class $\mathcal{KL}$ function candidate of $\beta(\cdot)$ in (6.35) *where $\mathcal{L} > 1$ is a constant.*

PROOF. For the notational simplicity in the sequent proof, we assume that $\mathcal{V}$ is complete; i.e., $\mathcal{N}_i = \mathcal{V} \setminus \{i\}$. If $(i,j) \notin \mathcal{E}$, then $\gamma_{ij}(s) = s$. We divide the remaining of the proof into three claims.

Claim E: The following hold for $i \in S_\ell \triangleq \{1, \cdots, \ell\}$:

$$\begin{aligned}
\|x_i\|_{[t_0,T]} \leq \max\{&\beta_i(\|x_i(t_0)\|, 0), \\
&\max_{\substack{(i,i_1,\cdots,i_\kappa)\in\mathcal{P}_{ii_\kappa} \\ i_1,\cdots,i_\kappa\in S_\ell}} \gamma_{ii_1} \circ \cdots \circ \gamma_{i_{\kappa-1}i_\kappa} \circ \gamma_{i_\kappa d}(\|d_{i_\kappa}\|_{[t_0,T]}), \\
&\max_{j\in S_\ell\setminus\{i\}} \max_{\substack{(j,i_\kappa,\cdots,i)\in\mathcal{P}_{ji} \\ i_1,\cdots,i_\kappa\in S_\ell}} \gamma_{ii_1} \circ \gamma_{i_1i_2} \circ \cdots \circ \gamma_{i_\kappa j} \circ \beta_j(\|x_j(t_0)\|, 0), \\
&\max_{j\notin S_\ell} \max_{\substack{(i,i_1,\cdots,i_\kappa,j)\in\mathcal{P}_{ij} \\ i_1,\cdots,i_\kappa\in S_\ell}} \gamma_{ii_1} \circ \cdots \circ \gamma_{i_\kappa j}(\|x_j\|_{[t_0,T]})\}. \qquad (6.36)
\end{aligned}$$

PROOF. By (6.34), one can see that

$$\|x_1\|_{[t_0,T]} \leq \max\{\beta_1(\|x_1(t_0)\|, 0), \gamma_{1d}(\|d_1\|_{[t_0,T]}), \max_{j\neq 1}\{\gamma_{1j}(\|x_j\|_{[t_0,T]})\}\}, \qquad (6.37)$$

and

$$\|x_2\|_{[t_0,T]} \leq \max\{\beta_2(\|x_2(t_0)\|, 0), \gamma_{2d}(\|d_2\|_{[t_0,T]}), \max_{j\neq 2}\{\gamma_{2j}(\|x_j\|_{[t_0,T]})\}\}. \qquad (6.38)$$

Substitute (6.38) into (6.37), and it renders the following:

$$\begin{aligned}
\|x_1\|_{[t_0,T]} \leq \max\{&\beta_1(\|x_1(t_0)\|, 0), \gamma_{1d}(\|d_1\|_{[t_0,T]}), \\
&\gamma_{12} \circ \beta_2(\|x_2(t_0)\|, 0), \gamma_{12} \circ \gamma_{2d}(\|d_2\|_{[t_0,T]}), \\
&\max_{j\neq 2}\{\gamma_{12} \circ \gamma_{2j}(\|x_j\|_{[t_0,T]})\}, \max_{j\notin\{1,2\}}\{\gamma_{1j}(\|x_j\|_{[t_0,T]})\}\}. \qquad (6.39)
\end{aligned}$$

Since $\gamma_{12} \circ \gamma_{21}$ is a contraction mapping, it follows from (6.39) that

$$\begin{aligned}
\|x_1\|_{[t_0,T]} \leq \max\{&\beta_1(\|x_1(t_0)\|, 0), \gamma_{1d}(\|d_1\|_{[t_0,T]}), \\
&\gamma_{12} \circ \beta_2(\|x_2(t_0)\|, 0), \gamma_{12} \circ \gamma_{2d}(\|d_2\|_{[t_0,T]}), \\
&\max_{j\notin\{1,2\}}\{\max\{\gamma_{1j}, \gamma_{12} \circ \gamma_{2j}\}(\|x_j\|_{[t_0,T]})\}\}. \qquad (6.40)
\end{aligned}$$

By symmetry, one can show a similar property to (6.40) for $\|x_2\|_{[t_0,T]}$. So (6.36) holds for the case of $\ell = 2$. Now assume that (6.36) holds for some $\ell < n$. Similar to (6.37), we have

$$\|x_{\ell+1}\|_{[t_0,T]} \leq \max\{\beta_{\ell+1}(\|x_{\ell+1}(t_0)\|, 0),$$

$$\gamma_{(\ell+1)d}(\|d_{\ell+1}\|_{[t_0,T]}), \max_{j\neq(\ell+1)}\{\gamma_{(\ell+1)j}(\|x_j\|_{[t_0,T]})\}\}. \tag{6.41}$$

Following analogous steps above, one can show that (6.37) holds for $\ell+1$. By induction, we complete the proof. ■

Claim F: The solution to (6.33) exists and it is bounded.

PROOF. A direct result of Claim E is that the following holds for all $i \in \mathcal{V}$:

$$\begin{aligned}\|x_i\|_{[t_0,T]} \leq \max\{&\beta_i(\|x_i(t_0)\|,0),\\ &\max_{\substack{(i,i_1,\cdots,i_\kappa)\in\mathcal{P}_{ii_\kappa}\\ i_1,\cdots,i_\kappa\in\mathcal{V}}} \gamma_{ii_1}\circ\cdots\circ\gamma_{i_{\kappa-1}i_\kappa}\circ\gamma_{i_\kappa d}(\|d_{i_\kappa}\|_{[t_0,T]}),\\ &\max_{j\neq i}\max_{\substack{(j,i_\kappa,\cdots,i)\in\mathcal{P}_{ji}\\ i_1,\cdots,i_\kappa\in\mathcal{V}}} \gamma_{ii_1}\circ\gamma_{i_1i_2}\circ\cdots\circ\gamma_{i_\kappa j}\circ\beta_j(\|x_j(t_0)\|,0)\}.\end{aligned} \tag{6.42}$$

Since all the gain functions γ_{ij} are contraction mappings, (6.42) renders the following:

$$\begin{aligned}\|x_i\|_{[t_0,T]} \leq \max\{&\beta_i(\|x_i(t_0)\|,0),\\ &\max_{j\in\mathcal{N}_i}\gamma_{ij}\circ\gamma_{jd}(\|d_j\|_{[t_0,T]}), \max_{j\in\mathcal{N}_i}\beta_j(\|x_j(t_0)\|,0)\}.\end{aligned} \tag{6.43}$$

Because of the choice of $x_i(t_0)$ and the bound on d, the relation (6.42) holds for any T. It implies that

$$\|x_i(t)\| \leq \max\{\beta_i(\|x_i(t_0)\|,0), \hat{\Delta}_d, \max_{j\in\mathcal{N}_i}\beta_j(\|x_j(t_0)\|,0)\}.$$

for all $t \geq t_0$ and thus is uniformly bounded. It completes the proof. ■

Claim G: System (6.33) is ISS; i.e., the following holds for all $i \in S_\ell \triangleq \{1,\cdots,\ell\}$:

$$\begin{aligned}\|x_i(t)\| \leq \max\{&\tilde{\beta}_i^{[\ell-1]}(\|x\|_\infty, t-t_0), \gamma_i^{[\ell-1]}(\|d\|_{[t_0,t]}),\\ &\max_{j\notin S_\ell}\max_{\substack{(i,i_1,\cdots,i_\kappa,j)\in\mathcal{P}_{ij}\\ i_1,\cdots,i_\kappa\in S_\ell}} \gamma_{ii_1}\circ\cdots\circ\gamma_{i_\kappa j}(\|x_j\|_{[t_0,t]})\},\end{aligned} \tag{6.44}$$

for some class $\mathcal{KL}$ function $\tilde{\beta}_i^{[\ell-1]}$ where $\|x\|_\infty \triangleq \sup\{\|x(t)\| \mid t \in [t_0,\infty]\}$.

PROOF. Let $\ell = 2$. Note that for any constant $\mathcal{L} > 1$,

$$\begin{aligned}
&\|x_1(t_0+T)\| \le \max\{\beta_1(\|x_1(t_0+\frac{2\mathcal{L}-1}{2\mathcal{L}}T)\|, \frac{1}{2\mathcal{L}}T),\\
&\gamma_{1d}(\|d_1\|_{[t_0+\frac{2\mathcal{L}-1}{2\mathcal{L}}T,t_0+T]}), \max_{j\neq 1}\gamma_{1j}(\|x_j\|_{[t_0+\frac{2\mathcal{L}-1}{2\mathcal{L}}T,t_0+T]})\}\\
&\le \max\{\beta_1(\|x\|_\infty, \frac{1}{2\mathcal{L}}T), \gamma_{1d}(\|d_1\|_{[t_0+\frac{2\mathcal{L}-1}{2\mathcal{L}}T,t_0+T]}),\\
&\max_{j\neq 1}\{\gamma_{1j}(\|x_j\|_{[t_0+\frac{2\mathcal{L}-1}{2\mathcal{L}}T,t_0+T]})\}\}. \qquad (6.45)
\end{aligned}$$

For any $\tau_2 \in [\frac{2\mathcal{L}-1}{2\mathcal{L}}T, T]$, it holds that

$$\begin{aligned}
&\|x_2(t_0+\tau_2)\| \le \max\{\beta_2(\|x_2(t_0+\frac{2\mathcal{L}-2}{2\mathcal{L}}T)\|, \tau_2 - \frac{2\mathcal{L}-2}{2\mathcal{L}}T),\\
&\gamma_{2d}(\|d_2\|_{[t_0+\frac{2\mathcal{L}-2}{2\mathcal{L}}T,t_0+\tau]}), \max_{j\neq 2}\{\gamma_{2j}(\|x_j\|_{[t_0+\frac{2\mathcal{L}-2}{2\mathcal{L}}T,t_0+\tau]})\}\}\\
&\le \max\{\beta_2(\|x\|_\infty, \frac{1}{\mathcal{L}}T), \gamma_{2d}(\|d_2\|_{[t_0+\frac{2\mathcal{L}-2}{2\mathcal{L}}T,t_0+T]}), \max_{j\neq 2}\{\gamma_{2j}(\|x_j\|_{[t_0+\frac{2\mathcal{L}-2}{2\mathcal{L}}T,t_0+T]})\}\}. \qquad (6.46)
\end{aligned}$$

So (6.46) implies that

$$\begin{aligned}
\|x_2\|_{[t_0+\frac{2\mathcal{L}-1}{2\mathcal{L}}T,t_0+T]} &\le \max\{\beta_2(\|x\|_\infty, \frac{1}{\mathcal{L}}T), \gamma_{2d}(\|d_2\|_{[t_0+\frac{2\mathcal{L}-2}{2\mathcal{L}}T,t_0+T]}),\\
&\max_{j\neq 2}\{\gamma_{2j}(\|x_j\|_{[t_0+\frac{2\mathcal{L}-2}{2\mathcal{L}}T,t_0+T]})\}\}. \qquad (6.47)
\end{aligned}$$

Substitute (6.47) into (6.45), and we have

$$\begin{aligned}
&\|x_1(t_0+T)\| \le \max\{\beta_1(\|x\|_\infty, \frac{1}{2\mathcal{L}}T), \gamma_{1d}(\|d_1\|_{[t_0+\frac{2\mathcal{L}-2}{2\mathcal{L}}T,t_0+T]}),\\
&\gamma_{12}\circ\beta_2(\|x\|_\infty, \frac{1}{\mathcal{L}}T), \gamma_{12}\circ\gamma_{2d}(\|d_2\|_{[t_0+\frac{2\mathcal{L}-2}{2\mathcal{L}}T,t_0+T]}),\\
&\gamma_{12}\circ\gamma_{21}(\|x_1\|_{[t_0+\frac{2\mathcal{L}-2}{2\mathcal{L}}T,t_0+T]}), \max_{j\notin S_2}\max\{\gamma_{1j}, \gamma_{12}\circ\gamma_{2j}\}(\|x_j\|_{[t_0+\frac{2\mathcal{L}-2}{2\mathcal{L}}T,t_0+T]})\}. \qquad (6.48)
\end{aligned}$$

Since $\gamma_{12}\circ\gamma_{21}(\cdot)$ is a contraction mapping, there is class $\mathcal{KL}$ function $\tilde{\beta}_1$ such that

$$\begin{aligned}
\|x_1(t)\| &\le \max\{\tilde{\beta}_1(\|x\|_\infty, t-t_0), \gamma_{1d}(\|d_1\|_{[t_0,t]}), \gamma_{12}\circ\gamma_{2d}(\|d_2\|_{[t_0,t]}),\\
&\max_{j\notin S_2}\max\{\gamma_{1j}, \gamma_{12}\circ\gamma_{2j}\}(\|x_j\|_{[t_0,t]})\}. \qquad (6.49)
\end{aligned}$$

By symmetry, there is class $\mathcal{KL}$ function $\tilde{\beta}_2$ such that

$$\begin{aligned}\|x_2(t)\| \leq \max\{&\tilde{\beta}_2(\|x\|_\infty, t-t_0), \gamma_{2d}(\|d_2\|_{[t_0,t]}), \gamma_{21}\circ\gamma_{1d}(\|d_1\|_{[t_0,t]}),\\ &\max_{j\notin S_2}\max\{\gamma_{2j}, \gamma_{21}\circ\gamma_{1j}\}(\|x_j\|_{[t_0,t]})\}.\end{aligned} \tag{6.50}$$

Hence, we have shown that (6.44) holds for $\ell = 2$. Now assume (6.44) holds for some $\ell < n$. Recall that

$$\|x_{\ell+1}(t)\| \leq \max\{\beta_{\ell+1}(\|x_{\ell+1}(t_0)\|, t-t_0), \gamma_{\ell+1}(\|d_{\ell+1}\|_{[t_0,t]}), \max_{j\neq\ell+1}\gamma_{ij}(\|x_j\|_{[t_0,t]})\}. \tag{6.51}$$

By using similar arguments towards the case of $\ell = 2$, one can show (6.44) holds for $\ell + 1$. Now we proceed to find a relation between $\|x\|_\infty$ and $\|d\|_\infty$. Because $\|x_i(t_0)\| \leq \|x(t_0)\|$, note that

$$\|x_i\|_\infty \leq \max\{\beta_i(\|x(t_0)\|, 0), \gamma_{id}(\|d_i\|_{[t_0,t]}), \max_{j\neq i}\{\gamma_{ij}(\|x_j\|_\infty)\}\}.$$

Similar to (6.44), one can show by induction that there are class $\mathcal{K}$ functions ρ_i and ρ_{id} such that

$$\|x_i\|_\infty \leq \max\{\rho_i(\|x(t_0)\|), \rho_{id}(\|d_i\|_{[t_0,t]})\}. \tag{6.52}$$

The combination of (6.52) and (6.44) achieves the desired result. ■

Now proceed with the proof that function

$$\beta(x,t) = |\mathcal{V}|\sum_{i\in\mathcal{V}}\beta_i(|\mathcal{V}|\sum_{k\in\mathcal{V}}\beta_k(x,0), \frac{t}{(2\mathcal{L})^{|\mathcal{V}|-1}})$$

is a candidate of class $\mathcal{KL}$ function β in (6.35). We first find candidates of functions $\tilde{\beta}_i^{[\ell-1]}$ in (6.44) and ρ_i in (6.52) and then combine them together. Note that by substituting (6.47) into (6.45), we have equation (6.48). Consider class $\mathcal{KL}$ functions in equation (6.48):

$$\begin{aligned}\|x_1(t_0+T)\| &\leq \max\{\beta_1(\|x\|_\infty, \frac{1}{2\mathcal{L}}T), \gamma_{12}\circ\beta_2(\|x\|_\infty, \frac{1}{\mathcal{L}}T)\}\\ &\leq \max\{\beta_1(\|x\|_\infty, \frac{1}{2\mathcal{L}}T) + \beta_2(\|x\|_\infty, \frac{1}{2\mathcal{L}}T)\}.\end{aligned}$$

This implies that, in (6.49), $\tilde{\beta}_1(x,t) = \sum_{k=1}^{2} \beta_k(x, \frac{t}{2\mathcal{L}})$ is a $\mathcal{KL}$ function candidate. Likewise, in (6.44),

$$\tilde{\beta}_i^{[\ell-1]}(x,t) = \sum_{k=1}^{\ell} \beta_k(x, \frac{t}{(2\mathcal{L})^{\ell-1}}) \tag{6.53}$$

is a $\mathcal{KL}$ function candidate for $\forall i \in S_\ell$ because we conduct $\ell - 1$ times of the substitutions. In a similar way, one can show that, in (6.52),

$$\rho_i(x) = \sum_{k=1}^{\ell} \beta_k(x, 0) \tag{6.54}$$

is a class $\mathcal{K}$ function candidate for $\forall i \in S_\ell$. Now we proceed to find a relation between $\tilde{\beta}_i^{[\ell-1]}$ and ρ_i when $S_\ell = \mathcal{V}$. With equation (6.52),

$$\|x\|_\infty \leq \sum_{i \in \mathcal{V}} \|x_i\|_\infty \leq |\mathcal{V}| \max_{i \in \mathcal{V}} \{\rho_i(\|x(t_0)\|), \rho_{id}(\|d_i\|_{[t_0,t]})\}. \tag{6.55}$$

By combining (6.44) and (6.55),

$$\|x_i(t)\| \leq \max\{\tilde{\beta}_i^{[|\mathcal{V}|-1]}(|\mathcal{V}| \max_{k \in \mathcal{V}} \rho_k(\|x(t_0)\|), t - t_0), \gamma_i^{[\ell-1]}(\|d\|_{[t_0,t]})\}.$$

This implies that

$$\beta(x,t) = |\mathcal{V}| \max_{i \in \mathcal{V}} \tilde{\beta}_i^{[|\mathcal{V}|-1]}(|\mathcal{V}| \max_{k \in \mathcal{V}} \rho_k(x), t) \tag{6.56}$$

is one of the class $\mathcal{KL}$ function candidates. By applying (6.53) and (6.54) to (6.56), we have the result. ■

Remark 6.8.1 *If functions $\beta_i(\cdot)$ in (6.34) for $\forall i \in \mathcal{V}$ are $\beta_i(x,t) = a_i^{-p_i(t)} r_i(x)$, then $\beta(\cdot)$ in (6.35) is also in the same form: $\beta(x,t) = a^{-p(t)} r(x)$ where $a, a_i > 0$ are constants, $p(t), p_i(t)$ are increasing functions without bound and $r(x), r_i(x)$ are class $\mathcal{K}$ functions.* ■

Remark 6.8.1 indicates that if functions $\beta_i(\cdot)$ are exponential functions, then $\beta(\cdot)$ is also an exponential function.

Chapter 7
Conclusion and future work

7.1 Conclusion

The objective of this dissertation is to design practical distributed controllers/mechanisms for CPS. We have investigated three emerging problems: cyber-physical security, cyber-physical economics, and smart grid. In particular, we have designed a multi-mode algorithm, where each mode is associated with a state and attack vector estimator, and the mode estimator chooses the most likely mode to generate the estimates of system states and attacker vectors. Attack-resilient machine learning algorithm has been proposed by incorporating attack-resilient state, attack vector and mode estimator with Gaussian process regression for a class of partially unknown nonlinear systems. We have proposed a bi-level lottery having user's heterogeneity parameter and social planner's perturbation parameter, and formulate an optimal bi-level lottery design problem where a Nash equilibrium of a lottery game is coincident with a socially optimal payoff or a greater payoff with least reward and perturbations. A convex approximation of the optimal bi-level lottery design problem is presented, where the approximation is exact under mild sufficient conditions. We have also proposed distributed controllers in smart power grid, with the integration of demand response. The proposed distributed controllers regulate angular frequencies to a desired common constant, rejecting the impact of unknown net loads.

7.2 Future work

In this section, we discuss possible future works and extensions, as follows.

1. **Distributed attack-resilient estimation.** We consider an extension of the NISME presented in Chapter 2 to distributed attack-resilient estimation, integrating consensus algorithms. The idea is motived by distributed Kalman filter for sensor networks [20, 172], where multiple sensors estimate states of the system in a distributed way. Distributed Kalman filter consists of a bank of Kalman filters and they share the outputs/covariance matrices via consensus algorithms, so that all local Kalman filters have the same estimates of states and covariance matrices. Likewise, the NISME can be distributed with embedding consensus algorithms. Distributed attack-resilient estimation would provide scalability of the algorithm and requires small communication bandwidths because it only requires local communication for consensus. Moreover, it is expected that distributed attack-resilient estimation is resistant to attacks targeting the estimator, because neighboring estimators would detect attacks when some of them fail. One limitation of distributed Kalman filter is that local Kalman filters estimate the network-wide states. This is computationally expensive and also needs network-wide information. There is an attempt to partition the network [173], so that local Kalman filers estimate partial states. Since it is realistic for a local system to have limited information and communication, our distributed attack-resilient estimation would base on partitioned distributed Kalman filter introduced in [173]. There are several challenges to address: (1) attack vectors might change faster before local authorities consent to their estimates; (2) shared covariance matrices are subject to local linearization points for nonlinear systems.

2. **Distributed attack-resilient control.** We have studied attack-resilient estimation in Chapters 2 and 3. It would be interesting to study *attack response*; i.e., attack-resilient distributed control. We consider attacks targeting communication channels between local control authorities and actuators such as replay attacks and denial-of-service attacks. The attacker hinders actuators from receiving generated control commands, and eventually ac-

tuators execute compromised control commands. We propose to adopt the methodology of model predictive control (MPC) [174]. A local control authority generates a sequence of control signals for several future steps. If no attack occurs, actuators execute newly generated control commands, and store future commands in local memories. If the system is under attack, the stored control commands are executed instead. The above idea has been applied to centralized control for linear systems [69] and distributed control for second-order mobile robots [68], where local systems are decoupled. When attack-resilient control is applied to distributed coupled systems, we face the challenges that (1) physical relations result in a cooperative game/non-cooperative game (2) stored control commands may destabilize the systems. To address the issues, the system operator desires to design a robust MPC with a mechanism, where the mechanism is supposed to help neighbors when they are under attacks, maintaining the network-wide performance.

3. **Attack-resilient machine learning.** Stability analysis of attack-resilient machine learning in Chapter 4 is performed by breaking the coupling between state estimation errors and function approximation errors. State estimation errors are independent from function approximation errors as analyzed in Theorem 4.5.1, while function approximation errors are subject to estimation errors in Theorem 4.5.2. We address the limitations of the current design and analysis as follows. Decoupling the errors induces Assumption 4.5.1. Assumption 4.5.1 is a sufficient condition of that used in the unknown input and output estimation for fully-known linear systems (Theorem 5 in [89]). The assumption may be restrictive in some cases, although it may be true for some CPS such as mobile robots. The assumption would be relaxed if two stability analysis are combined.

4. **Attack-resilient estimation.** In Chapter 2, we consider a class of systems subject to additive noises. Since Cyber-physical systems connect many heterogeneous systems, it would be practical to extend the existing estimation algorithm to robust to multiplicative and additive noises. The idea of particle filter and Unscented Kalman filter could be used, because they are robust extension of Kalman filter. The both filters use multiple testing points to compute accurate posterior distributions of the internal states, while Kalman

filter use only one testing point (mean).

5. **Attack-resilient estimation for nonlinear systems.** Attack-resilient estimation is an emerging research area and most of the literature consider linear systems. To apply existing algorithms to nonlinear systems, it is required to linearize the target systems. The proposed tool is the feedback linearizion, where the nonlinear system is transformed to equivalent linear system with the change of variables. To apply the feedback linearization, one needs to investigate whether linearization points/transformation are approximated, and investigate how the feedback linearization errors affect the performance.

Bibliography

[1] X. Zhang, C. Rehtanz, and B. Pal. *Flexible AC transmission systems: modelling and control*. Springer Science & Business Media, 2012.

[2] C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*.

[3] R. Christie. Power systems test case archive. *Electrical Engineering dept., University of Washington*, 2000.

[4] J. P. Holdren, E. Lander, and H. Varmus. Report to the president and congress: Designing a digital future: federally funded research and development in networking and information technology. *Executive Office of the President and Presidents Council of Advisors on Science and Technology*, 2010.

[5] F. Bullo, J. Cortes, and S. Martinez. *Distributed control of robotic networks: a mathematical approach to motion coordination algorithms*. Princeton University Press, 2009.

[6] M. Mesbahi and M. Egerstedt. *Graph theoretic methods in multiagent networks*. Princeton University Press, 2010.

[7] W. Ren and R. W. Beard. *Distributed consensus in multi-vehicle cooperative control*. Springer, 2008.

[8] M. Zhu and S. Martínez. *Distributed optimization-based control of multi-agent networks in complex environments*. Springer, 2015.

[9] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and distributed computation: numerical methods*. Prentice hall Englewood Cliffs, NJ, 1989.

[10] L. S. Lasdon. *Optimization theory for large systems*. Courier Corporation, 1970.

[11] J. N. Tsitsiklis. *Problems in decentralized decision making and computation*. Massachusetts Inst of Tech Cambridge Lab for Information and Decision Systems, 1984.

[12] T. Başar and G. J. Olsder. *Dynamic noncooperative game theory*. SIAM, 1998.

[13] K. J. Arrow and G. Debreu. Existence of an equilibrium for a competitive economy. *Econometrica: Journal of the Econometric Society*, pages 265–290, 1954.

[14] F. Dorfler and F. Bullo. Synchronization and transient stability in power networks and nonuniform kuramoto oscillators. *SIAM Journal on Control and Optimization*, 50(3):1616–1642, 2012.

[15] N. Li, L. Chen, and S. H. Low. Optimal demand response based on utility maximization in power networks. In *IEEE Power and Energy Society General Meeting*, pages 1–8. IEEE, 2011.

[16] J. R. T. Lawton, R. W. Beard, and B. J. Young. A decentralized approach to formation maneuvers. *IEEE Transactions on Robotics and Automation*, 19(6):933–941, 2003.

[17] Z. Lin, B. Francis, and M. Maggiore. Necessary and sufficient graphical conditions for formation control of unicycles. *IEEE Transactions on Automatic Control*, 50(1):121–127, 2005.

[18] X. Wang, V. Yadav, and S. N. Balakrishnan. Cooperative UAV formation flying with obstacle/collision avoidance. *IEEE Transactions on Control Systems Technology*, 15(4):672–679, 2007.

[19] M. Rabbat and R. Nowak. Distributed optimization in sensor networks. In *Proceedings of International Symposium on Information Processing in Sensor Networks*, pages 20–27. ACM, 2004.

[20] R. Olfati-Saber. Distributed kalman filtering for sensor networks. In *IEEE Conference on Decision and Control*, pages 5492–5498. IEEE, 2007.

[21] Z. Wang, L. Wang, A. I. Dounis, and R. Yang. Multi-agent control system with information fusion based comfort model for smart buildings. *Applied Energy*, 99:247–254, 2012.

[22] P. Lichodzijewski, A. N. Zincir-Heywood, and M. I. Heywood. Host-based intrusion detection using self-organizing maps. In *Proceedings of International Joint Conference on Neural Networks*, pages 1714–1719. IEEE, 2002.

[23] S. Mukkamala, G. Janoski, and A. Sung. Intrusion detection using neural networks and support vector machines. In *Proceedings of International Joint Conference on Neural Networks*, pages 1702–1707. IEEE, 2002.

[24] S. Axelsson. The base-rate fallacy and the difficulty of intrusion detection. *ACM Transactions on Information and System Security*, 3(3):186–205, 2000.

[25] W. Lee and S. J. Stolfo. Data mining approaches for intrusion detection. In *USENIX Security Symposium*, pages 79–93, 1998.

[26] M. Roesch. Snort: Lightweight intrusion detection for networks. In *Lisa*, pages 229–238, 1999.

[27] K. Zhou and J. C. Doyle. *Essentials of robust control*. Prentice hall Upper Saddle River, NJ, 1998.

[28] M. Krstic, P. V. Kokotovic, and I. Kanellakopoulos. *Nonlinear and adaptive control design*. John Wiley & Sons, Inc., 1995.

[29] K. J. Åström. *Introduction to stochastic control theory*. Courier Corporation, 2012.

[30] Y. Liu, P. Ning, and M. K. Reiter. False data injection attacks against state estimation in electric power grids. *ACM Transactions on Information and System Security*, 14(1):13, 2011.

[31] F. Pasqualetti, F. Dörfler, and F. Bullo. Attack detection and identification in cyber-physical systems. *IEEE Transactions on Automatic Control*, 58(11):2715–2729, 2013.

[32] H. Fawzi, P. Tabuada, and S. Diggavi. Secure estimation and control for cyber-physical systems under adversarial attacks. *IEEE Transactions on Automatic Control*, 59(6):1454–1467, 2014.

[33] M. Pajic, J. Weimer, N. Bezzo, P. Tabuada, O. Sokolsky, I. Lee, and G. J. Pappas. Robustness of attack-resilient state estimators. In *International Conference on Cyber-Physical Systems*, pages 163–174, 2014.

[34] Y. Mo and B. Sinopoli. False data injection attacks in control systems. In *Workshop on Secure Control Systems*, pages 1–6, 2010.

[35] Y. Shoukry, M. Chong, M. Wakaiki, P. Nuzzo, A. L. Sangiovanni-Vincentelli, S. A. Seshia, J. P. Hespanha, and P. Tabuada. SMT-based observer design for cyber-physical systems under sensor attacks. In *Proceedings of the International Conference on Cyber-Physical Systems*, pages 1–10, 2016.

[36] S. Z. Yong, M. Zhu, and E. Frazzoli. Resilient state estimation against switching attacks on stochastic cyber-physical systems. In *IEEE Conference on Decision and Control*, pages 5162–5169, 2015.

[37] D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

[38] S. A. Goldman and R. H. Sloan. Can PAC learning algorithms tolerate random attribute noise? *Algorithmica*, 14(1):70–84, 1995.

[39] L. G. Valiant. Learning disjunction of conjunctions. In *IJCAI*, pages 560–566, 1985.

[40] M. Kearns and M. Li. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22(4):807–837, 1993.

[41] B. Safarinejadian and E. Kowsari. Fault detection in non-linear systems based on GP-EKF and GP-UKF algorithms. *Systems Science & Control Engineering: An Open Access Journal*, 2(1):610–620, 2014.

[42] J. Ko, D. J. Kleint, D. Fox, and D. Haehnelt. GP-UKF: Unscented kalman filters with gaussian process prediction and observation models. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1901–1907. IEEE, 2007.

[43] E. Kowsari, B. Safarinejadian, and J. Zarei. Non-parametric fault detection methods in non-linear systems. *IET Science, Measurement & Technology*, 10(3):167–176, 2016.

[44] E. Clarke. Multipart pricing of public goods. *Public Choice*, 11(1):17–33, 1971.

[45] T. Groves. Incentives in teams. *Econometrica*, 41(4):617–631, 1973.

[46] W. Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of Finance*, 16(1):8–37, 1961.

[47] J. Green and J. J. Laffont. Characterization of satisfactory mechanisms for the revelation of preferences for public goods. *Econometrica: Journal of the Econometric Society*, pages 427–438, 1977.

[48] B. Holmström. Groves' scheme on restricted domains. *Econometrica: Journal of the Econometric Society*, pages 1137–1144, 1979.

[49] P. Bolton and M. Dewatripont. *Contract theory*. MIT press, 2005.

[50] S. Shavell. Risk sharing and incentives in the principal and agent relationship. *The Bell Journal of Economics*, pages 55–73, 1979.

[51] D. P. Baron and R. B. Myerson. Regulating a monopolist with unknown costs. *Econometrica: Journal of the Econometric Society*, pages 911–930, 1982.

[52] M. Spence. Job market signaling. *The quarterly journal of Economics*, 87(3):355–374, 1973.

[53] J. A. Mirrlees. An exploration in the theory of optimum income taxation. *The review of economic studies*, 38(2):175–208, 1971.

[54] R. B. Myerson and M. A. Satterthwaite. Efficient mechanisms for bilateral trading. *Journal of economic theory*, 29(2):265–281, 1983.

[55] N. Nisan and A. Ronen. Algorithmic mechanism design. In *Annual ACM Symposium on Theory of computing*, pages 129–140, 1999.

[56] A. Petcu, B. Faltings, and D.C. Parkes. MDPOP: Faithful distributed implementation of efficient social choice problems. In *International Joint Conference on Autonomous Agents and Multi-agent Systems*, pages 1397–1404, 2006.

[57] T. Tanaka, F. Farokhi, and C. Langbort. A faithful distributed implementation of dual decomposition and average consensus algorithms. In *IEEE Conference on Decision and Control*, pages 2985–2990, 2013.

[58] T. Basar. Affine incentive schemes for stochastic systems with dynamic information. *SIAM Journal on Control and Optimization*, 22(2):199–210, 1984.

[59] Y. Ho, P. Luh, and G. Olsder. A control-theoretic view on incentives. *Automatica*, 18(2):167–179, 1982.

[60] J. Morgan. Financing public goods by means of lotteries. *Review of Economic Studies*, 67:761–784, 2000.

[61] Renewables REN21. Global status report, Paris, REN21 Secretariat, 2014.

[62] K. Y. Lim, Y. Wang, and R. Zhou. Robust decentralised load-frequency control of multi-area power systems. 143(5):377–386, 1996.

[63] H. Bevrani, Y. Mitani, K. Tsuji, and H. Bevrani. Bilateral based robust load frequency control. *Energy Conversion and Management*, 46(7):1129–1146, 2005.

[64] H. Shayeghi. A robust decentralized power system load frequency control. *Journal of Electrical Engineering*, 59(6):281–293, 2008.

[65] S. Trip, M. Bürger, and C. De Persis. An internal model approach to (optimal) frequency regulation in power grids with time-varying voltages. *Automatica*, 64:240–253, 2016.

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

[67] J. Slay and M. Miller. Lessons learned from the maroochy water breach. In *International Conference on Critical Infrastructure Protection*, pages 73–82. Springer, 2007.

[68] M. Zhu and S. Martínez. On distributed constrained formation control in operator–vehicle adversarial networks. *Automatica*, 49(12):3571–3582, 2013.

[69] M. Zhu and S. Martínez. On the performance analysis of resilient networked control systems under replay attacks. *IEEE Transactions on Automatic Control*, 59(3):804–808, 2014.

[70] J. Weimer, S. Kar, and K. H. Johansson. Distributed detection and isolation of topology attacks in power networks. In *Proceedings of International Conference on High Confidence Networked Systems*, pages 65–72. ACM, 2012.

[71] M. Darouach and M. Zasadzinski. Unbiased minimum variance estimation for systems with unknown exogenous inputs. *Automatica*, 33(4):717–719, 1997.

[72] M. Hou and R. J. Patton. Optimal filtering for systems with unknown inputs. *IEEE Transactions on Automatic Control*, 43(3):445–449, 1998.

[73] C. Hsieh. Robust two-stage kalman filters for systems with unknown inputs. *IEEE Transactions on Automatic Control*, 45(12):2374–2378, 2000.

[74] S. Gillijns and B. De Moor. Unbiased minimum-variance input and state estimation for linear discrete-time systems with direct feedthrough. *Automatica*, 43(5):934–937, 2007.

[75] S. Z. Yong, M. Zhu, and E. Frazzoli. Simultaneous input and state estimation for linear discrete-time stochastic systems with direct feedthrough. In *IEEE Conference on Decision and Control*, pages 7034–7039, 2013.

[76] S. Z. Yong, M. Zhu, and E. Frazzoli. A unified filter for simultaneous input and state estimation of linear discrete-time stochastic systems. *Automatica*, 63(1):321–329, 2016.

[77] S. J. Liu, J. F. Zhang, and Z. P. Jiang. A notion of stochastic input-to-state stability and its application to stability of cascaded stochastic nonlinear systems. *Acta Mathematicae Applicatae Sinica, English Series*, 24(1):141–156, 2008.

[78] M. Krstic and H. Deng. *Stabilization of nonlinear uncertain systems*. Springer-Verlag New York, Inc., 1998.

[79] B. D. O. Anderson and J. B. Moore. Detectability and stabilizability of time-varying discrete-time linear systems. *SIAM Journal on Control and Optimization*, 19(1):20–32, 1981.

[80] A. J. Wood and B. F. Wollenberg. *Power Generation Operation and Control*. New York: Wiley, 1996.

[81] J. De La Ree, V. Centeno, J. S. Thorp, and A. G. Phadke. Synchronized phasor measurement applications in power systems. *IEEE Transactions on Smart Grid*, 1(1):20–27, 2010.

[82] A. G. Phadke. Synchronized phasor measurements in power systems. *IEEE Computer Applications in Power*, 6(2):10–15, 1993.

[83] H. K. Alfares and M. Nazeeruddin. Electric load forecasting: Literature survey and classification of methods. *International Journal of Systems Science*, 33(1):23–34, 2002.

[84] H. S. Hippert, C. E. Pedreira, and R. C. Souza. Neural networks for short-term load forecasting: A review and evaluation. *IEEE Transactions on power systems*, 16(1):44–55, 2001.

[85] F.L. Alvarado, J. Meng, C.L. DeMarco, and W.S. Mota. Stability analysis of interconnected power systems coupled with market dynamics. *IEEE Transactions on power systems*, 16(4):695–701, 2001.

[86] M. K. von Renesse and M. Scheutzow. Existence and uniqueness of solutions of stochastic functional differential equations. *Random Operators and Stochastic Equations*, 18(3):267–284, 2010.

[87] J. H. Kotecha and P. M. Djuric. Gaussian sym particle filtering. *IEEE Transactions on Signal Processing*, 51(10):2592–2601, 2003.

[88] A. Papoulis and S. U. Pillai. *Probability, random variables, and stochastic processes*. Tata McGraw-Hill Education, 2002.

[89] S. Z. Yong, M. Zhu, and E. Frazzoli. A unified filter for simultaneous input and state estimation of linear discrete-time stochastic systems. *Automatica*, 63:321–329, 2016.

[90] S. Z. Yong, M. Zhu, and E. Frazzoli. Simultaneous input and state estimation for linear time-varying continuous-time stochastic systems. *IEEE Transactions on Automatic Control*, 62(5):2531–2538, 2017.

[91] S. Z. Yong, M. Zhu, and E. Frazzoli. Switching and data injection attacks on stochastic cyber-physical systems: Modeling, resilient estimation, and attack mitigation. *ACM Transactions on Cyber-Physical Systems*, 2(2):9, 2018.

[92] M. Boutayeb, H. Rafaralahy, and M. Darouach. Convergence analysis of the extended kalman filter used as an observer for nonlinear deterministic discrete-time systems. *IEEE Transactions on Automatic Control*, 42(4):581–586, 1997.

[93] S. Kluge, K. Reif, and M. Brokate. Stochastic stability of the extended Kalman filter with intermittent observations. *IEEE Transactions on Automatic Control*, 55(2):514–518, 2010.

[94] K. Reif, S. Günther, E. Yaz, and R. Unbehauen. Stochastic stability of the discrete-time extended kalman filter. *IEEE Transactions on Automatic Control*, 44(4):714–728, 1999.

[95] Y. Song and J. W. Grizzle. The extended kalman filter as a local asymptotic observer for nonlinear discrete-time systems. In *American Control Conference, 1992*, pages 3365–3369, 1992.

[96] M. I. Ostrovskii. *Metric embeddings: bilipschitz and coarse embeddings into Banach spaces*. Walter de Gruyter, 2013.

[97] C. Truesdell and W. Noll. *The non-linear field theories of mechanics*. Springer, 2004.

[98] J. Luukkainen. Rings of functions in lipschitz topology. *Ann. Acad. Sci. Fenn. Ser. AI Math*, 4:119–135, 1978.

[99] A. H. Jazwinski. *Stochastic processes and filtering theory*. Courier Corporation, 2007.

[100] D. Simon. *Optimal state estimation: Kalman, H infinity, and nonlinear approaches*. John Wiley & Sons, 2006.

[101] W. W. Hager. Updating the inverse of a matrix. *SIAM review*, 31(2):221–239, 1989.

[102] J. Hoffman-Jorgensen. *Probability with a view towards statistics*. CRC Press, 1994.

[103] T. Kailath, A. H. Sayed, and B. Hassibi. *Linear estimation*. Prentice Hall Upper Saddle River, NJ, 2000.

[104] M. D. Ilic, L. Xie, U. A. Khan, and J. M. F. Moura. Modeling of future cyber–physical energy systems for distributed sensing and control. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 4(40):825–838, 2010.

[105] P. Kundur, N.J. Balu, and M.G. Lauby. *Power system stability and control*. McGraw-Hill, 1994.

[106] H. Kim, M. Zhu, and J. Lian. Distributed robust adaptive frequency regulation of power grid with dynamic loads. *arXiv preprint arXiv:1510̇05071*, 2017. `http://arxiv.org/abs/1510.05071`.

[107] K-team mobile robotics - Khepera III. `http://www.k-team.com/mobile-robotics-products/old-products/khepera-iii`, 2016.

[108] K-team mobile robotics - Korebot II. `http://www.k-team.com/mobile-robotics-products/old-products/korebot-ii`, 2016.

[109] M. H. Hebert, C. E. Thorpe, and A. Stentz. *Intelligent unmanned ground vehicles: autonomous navigation research at Carnegie Mellon*. Springer Science & Business Media, 2012.

[110] J. Leonard, J. How, S. Teller, M. Berger, S. Campbell, G. Fiore, L. Fletcher, E. Frazzoli, A. Huang, S. Karaman, et al. A perception-driven autonomous urban vehicle. *Journal of Field Robotics*, 2008.

[111] V. A. Bavdekar, A. P. Deshpande, and S. C. Patwardhan. Identification of process and measurement noise covariance for state and parameter estimation using extended kalman filter. *Journal of Process Control*, 2011.

[112] Boston Dynamics. Bigdog overview, 2008.

[113] C. Chen, A. Seff, A. Kornhauser, and J. Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *IEEE International Conference on Computer Vision*, pages 2722–2730. IEEE, 2015.

[114] M. Behl, A. Jain, and R. Mangharam. Data-driven modeling, control and tools for cyber-physical energy systems. In *ACM/IEEE International Conference on Cyber-Physical Systems*, pages 1–10. IEEE, 2016.

[115] Y. Zhang, M. Qiu, C. W. Tsai, M. M. Hassan, and A. Alamri. Health-CPS: Healthcare cyber-physical system assisted by cloud and big data. *IEEE Systems Journal*, 11(1):88–95, 2017.

[116] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[117] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2015.

[118] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. Practical black-box attacks against machine learning. In *Proceedings of ACM on Asia Conference on Computer and Communications Security*, pages 506–519. ACM, 2017.

[119] Y. Mo, E. Garone, A. Casavola, and B. Sinopoli. False data injection attacks against state estimation in wireless sensor networks. In *IEEE Conference on Decision and Control*, pages 5967–5972. IEEE, 2010.

[120] D. J. C. MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[121] E. Snelson and Z. Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems*, pages 1257–1264, 2006.

[122] P. Whittle. The risk-sensitive certainty equivalence principle. *Journal of Applied Probability*, pages 383–388, 1986.

[123] B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *International Conference on Computational Learning Theory*, pages 416–426. Springer, 2001.

[124] J. Ko and D. Fox. GP-Bayesfilters: Bayesian filtering using Gaussian process prediction and observation models. *Autonomous Robots*, 27(1):75–90, 2009.

[125] C. Pluntke and B. Prabhakar. Insinc: A platform for managing peak demand in public transit. *JOURNEYS, Land Transport Authority Academy of Singapore*, pages 31–39, 2013.

[126] D. Merugu, B. Prabhakar, and N. Rama. An incentive mechanism for decongesting the roads: A pilot program in bangalore. In *ACM NetEcon Workshop*, 2009.

[127] P. N. Barbieri, M. Bigoni, and M. Fabbri. Incentives against free-riding: A field experiment final report. 2015.

[128] J. S. Laguilles, E. A. Williams, and D. B. Saunders. Can lottery incentives boost web survey response rates? findings from four experiments. *Research in Higher Education*, 52(5):537–553, 2011.

[129] J. Robertson, E. J. Walkom, and P. McGettigan. Response rates and representativeness: a lottery incentive improves physician survey return rates. *Pharmacoepidemiology and Drug Safety*, 14(8):571–577, 2005.

[130] P. V. Sundar Balakrishnan, S. K. Chawla, M. F. Smith, and B. P. Michalski. Mail survey response rates using a lottery prize giveaway incentive. *Journal of Direct Marketing*, 6(3):54–59, 1992.

[131] J. Li, B. Xia, X. Geng, H. Ming, S. Shakkottai, V. Subramanian, et al. Energy coupon: A mean field game perspective on demand response in smart grids. *ACM SIGMETRICS Performance Evaluation Review*, 43(1):455–456, 2015.

[132] G. A. Schwartz, H. Tembine, S. Amin, and S. S. Sastry. Demand response scheme based on lottery-like rebates. *IFAC Proceedings Volumes*, 47(3):4584–4588, 2014.

[133] B. Djehiche, A. Tcheukam, and H. Tembine. Mean-field-type games in engineering. *arXiv preprint arXiv:1605.03281*, 2016.

[134] P. Loiseau, G. Schwartz, J. Musacchio, S. Amin, and S. S. Sastry. Congestion pricing using a raffle-based scheme. In *International Conference on Network Games, Control and Optimization*, pages 1–8. IEEE, 2011.

[135] A. Lange, J. A. List, and M. K. Price. Using lotteries to finance public goods: Theory and experimental evidence. *International Economic Review*, 48(3):901–927, 2007.

[136] D. S. Damianov and R. Peeters. On the disclosure of ticket sales in charitable lotteries. *Economics Letters*, 150:73–76, 2017.

[137] P. Pecorino and A. Temimi. Lotteries, group size, and public good provision. *Journal of Public Economic Theory*, 9(3):451–465, 2007.

[138] J. Nash. Non-cooperative games. *Annals of mathematics*, pages 286–295, 1951.

[139] E. Koutsoupias and C. Papadimitriou. Worst-case equilibria. *STACS 99*, pages 404–413, 1999.

[140] S. G. Krantz and H. R. Parks. *The implicit function theorem: history, theory, and applications*. Springer Science & Business Media, 2012.

[141] E. Anshelevich, A. Dasgupta, J. Kleinberg, E. Tardos, T. Wexler, and T. Roughgarden. The price of stability for network design with fair cost allocation. *SIAM Journal on Computing*, 38(4):1602–1623, 2008.

[142] J. F. Bard. Some properties of the bilevel programming problem. *Journal of Optimization Theory and Applications*, 68(2):371–378, 1991.

[143] R. G. Jeroslow. The polynomial hierarchy and a simple model for competitive analysis. *Mathematical Programming*, 32(2):146–164, 1985.

[144] L. Vicente, G. Savard, and J. Júdice. Descent approaches for quadratic bilevel programming. *Journal of Optimization Theory and Applications*, 81(2):379–399, 1994.

[145] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas. MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education. *IEEE Transactions on Power Systems*, 26(1):12–19, 2011.

[146] C. Cobb and P. Douglas. A theory of production. *The American Economic Review*, pages 139–165, 1928.

[147] R. McCleary and R. Barro. US-based private voluntary organizations: Religious and secular pvos engaged in international relief & development. Technical report, National Bureau of Economic Research, 2006.

[148] D. Ribar and M. Wilhelm. Altruistic and joy-of-giving motivations in charitable behavior. *Journal of Political Economy*, 110(2):425–457, 2002.

[149] J. Rotemberg. Charitable giving when altruism and similarity are linked. *Journal of Public Economics*, 114:36–49, 2014.

[150] M. Grant, S. Boyd, and Y. Ye. CVX: Matlab software for disciplined convex programming, 2008.

[151] F. Alvarado. The stability of power system markets. *IEEE Transactions on Power Systems*, 14(2):505–511, 1999.

[152] W. Chiu, H. Sun, and H. V. Poor. Energy imbalance management using a robust pricing scheme. *IEEE Transactions on Smart Grid*, 4(2):896–904, 2013.

[153] P. Milan, M. Wächter, and J. Peinke. Turbulent character of wind energy. *Physical Review Letters*, 110(13):138701, 2013.

[154] I. Van der Hoven. Power spectrum of horizontal wind speed in the frequency range from 0.0007 to 900 cycles per hour. *Journal of Meteorology*, 14(2):160–164, 1957.

[155] L. A. Aguirre, D. D. Rodrigues, S. T. Lima, and C. B. Martinez. Dynamical prediction and pattern mapping in short-term load forecasting. *International Journal of Electrical Power & Energy Systems*, 30(1):73–82, 2008.

[156] G. P. Tolstov. *Fourier series*. Courier Corporation, 2012.

[157] B. A. Francis and W. M. Wonham. The internal model principle for linear multivariable regulators. *Applied Mathematics and Optimization*, 2(2):170–194, 1975.

[158] A. Isidori and C. I. Byrnes. Output regulation of nonlinear systems. *IEEE Transactions on Automatic Control*, 35(2):131–140, 1990.

[159] B. A. Francis and W. M. Wonham. The internal model principle of control theory. *Automatica*, 12(5):457–465, 1976.

[160] R. Bhatia and P. Rosenthal. How and why to solve the operator equation ax- xb= y. *Bulletin of the London Mathematical Society*, 29(01):1–21, 1997.

[161] M. Krstic, I. Kanellakopoulos, and P. Kokotovic. *Nonlinear and Adaptive Control Design*. John Wiley and Sons, 1995.

[162] B. N. Parlett. The rayleigh quotient iteration and some generalizations for nonnormal matrices. *Mathematics of Computation*, 28(127):679–693, 1974.

[163] H. K. Khalil. *Nonlinear Systems*. Upper Saddle River: Prentice hall, 2002.

[164] E. Sontag. Smooth stabilization implies coprime factorization. *IEEE Transactions on Automatic Control*, 34(4):435–443, 1989.

[165] N. Srebro. *Learning with matrix factorizations*. PhD thesis, Massachusetts Institute of Technology, 2004.

[166] D. Trudnowski and J. Undrill. The MinniWECC system model. *Oscillation damping controls*, 2008.

[167] D. Trudnowski, D. Kosterev, and J. Undrill. PDCI damping control analysis for the western north american power system. In *Power and Energy Society General Meeting*, pages 1–5. IEEE, 2013.

[168] J. Lian, Q. Zhang, L. D. Marinovici, R. Fan, and J. Hansen. Wide-area demand-side control for inter-area oscillation mitigation in power systems. In *Transmission and Distribution Conference and Exposition*, pages 1–5. IEEE, 2018.

[169] B. Pal and B. Chaudhuri. *Robust control in power systems*. Springer Science & Business Media, 2006.

[170] G. Rogers. *Power system oscillations*. Springer Science & Business Media, 2012.

[171] M. Zhu and J. Huang. Small gain theorem with restrictions for uncertain time-varying nonlinear systems. *Communication in Information and Systems*, 6(2):115–136, 2006.

[172] R. Olfati-Saber. Distributed kalman filter with embedded consensus filters. In *IEEE Conference on Decision and Control*, pages 8179–8184. IEEE, 2005.

[173] U. A. Khan and J. Moura. Distributing the kalman filter for large-scale systems. *IEEE Transactions on Signal Processing*, 56(10):4919–4935, 2008.

[174] C. E. Garcia, D. M. Prett, and M. Morari. Model predictive control: theory and practice–a survey. *Automatica*, 25(3):335–348, 1989.

Vita

Hunmin Kim

Hunmin Kim is a post doctoral researcher in Mechanical Science and Engineering at University of Illinois at Urbana-Champaign. He received the Ph.D. degree in Electrical Engineering in 2018 from Pennsylvania State University. He got his B.S. degree in Mechanical Engineering in 2012 from Pusan National University and graduated second in class. His research focuses on a broad area of distributed control, estimation, and game theory for multi-agent systems.