

**Toward a More Reliable Power System:
Frequency Regulation from Buildings and Secure Estimation against
Cyber Attacks**

by

Qie Hu

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy
in
Engineering - Electrical Engineering and Computer Sciences
in the
Graduate Division
of the
University of California, Berkeley

Committee in charge:

Professor Claire Tomlin, Chair
Professor Murat Arcak
Professor Anil Aswani

Fall 2017

Toward a More Reliable Power System:
Frequency Regulation from Buildings and Secure Estimation against Cyber Attacks

Copyright © 2017

by

Qie Hu

Abstract

Toward a More Reliable Power System:
Frequency Regulation from Buildings and Secure Estimation against Cyber Attacks

by

Qie Hu

Doctor of Philosophy in Engineering - Electrical Engineering and Computer Sciences

University of California, Berkeley
Professor Claire Tomlin, Chair

This thesis presents progress in overcoming two challenges in achieving a reliable electric power system: frequency regulation and secure state estimation against cyber attacks.

Frequency regulation is a type of ancillary service used to control the grid frequency around its nominal value. Recently, the higher penetration of renewable energy sources has increased the demand for frequency regulation reserves. This thesis explores the feasibility of using commercial buildings for this application. Commercial buildings are a tremendous untapped source because of their large consumption and thermal inertia, as well as being able to adjust their electricity consumption continuously. However, large disturbances such as occupancy and the complexity of the heating, ventilation and air conditioning system of commercial buildings make it challenging to: 1) identify a model that accurately describes its temperature evolution and is amenable to control, 2) design a robust frequency regulation controller. This thesis tackles both challenges. First, it proposes a physics-based and a data-driven method to **identity** a building model **that is suitable with the building in regular operation and captures internal gains such as occupancy**. Both methods are used to identify models of the same testbed, and a quantitative comparison of the resulting models is made including open-loop prediction accuracy and closed-loop control performance. It is concluded that a data-driven model may be suitable for temperature critical applications such as frequency regulation. Second, this thesis improves on existing frequency regulation control schemes and proposes a bilevel controller that is suitable for buildings subject to larger uncertainties, where accurate models are unavailable. Finally, field experiments in accordance with the Pennsylvania, New Jersey and Maryland's regulation market rules are conducted on an occupied building during both daytime and nighttime, which demonstrate the suitability of the data-driven building model and the performance of the proposed frequency regulation controller.

On the other hand, instruments such as phasor measurement units that communicate over wireless networks to achieve better efficiency of the power system are becoming increasingly pervasive, especially under the smart grid initiatives. However,

these communication networks are vulnerable to cyber attacks that can be erratic and difficult to predict. To add to the challenge, the dynamics of the power system cannot be approximated by a linear model when it's under severe disturbances. This thesis first develops a secure state estimation method for linear dynamical systems under sensor attack, and then extends it to two classes of nonlinear systems and applies it to the nonlinear power system. Both estimation methods assume that the attack signal can be arbitrary and unbounded, and the set of attacked sensors can change over time. More specifically, we use feedback linearization to transform the nonlinear system into an equivalent linear system. We then formulate the secure estimation problem into a classical error correction problem, from which we propose an l_1 -optimization based estimator that is computationally efficient. In addition, we prove the maximum number of sensor attacks that can be corrected with our estimator and propose to use pole placement techniques to design a feedback controller such that the resulting secure estimator can guarantee accurate estimation. Finally, to improve the estimator's practical performance, we propose to combine our secure estimator with a KF, where the KF serves to filter out both occasional estimation attacks by the secure estimator and noisy measurements.

Contents

Contents	i
1 Introduction	1
1.1 Frequency Regulation from Commercial Buildings	1
1.2 Secure Estimation against Cyber Attacks	3
 1 Commercial Buildings for Frequency Regulation	 4
2 Mathematical Modeling of Commercial Buildings	5
2.1 Introduction	6
2.2 Preliminaries	7
2.3 Data-Driven Model	8
2.4 Physics-Based Model	13
2.5 Quantitative Comparison of both Models	17
2.6 Conclusion	23
 3 Experimental Demonstration of Frequency Regulation from Commercial Buildings	 25
3.1 Introduction	26
3.2 Problem Statement	27
3.3 Model Identification	30
3.4 Two-Level Control Scheme	31
3.5 Experimental Results	36
3.6 Conclusion	40
 2 Secure Estimation under Cyber Attacks	 44
 4 Secure Estimation for Linear Systems	 45
4.1 Introduction	46
4.2 Review of Classical Error Correction	47
4.3 Secure Estimation	48

4.4	Estimator Design	53
4.5	Numerical Examples	57
4.6	Conclusion	63
5	Secure Estimation for Nonlinear Power Systems	68
5.1	Introduction	69
5.2	Secure Estimation for Nonlinear Systems	70
5.3	Power System State Estimation	73
5.4	Secure State Estimator for Wide Area Control Systems	77
5.5	Numerical Example	83
5.6	Conclusion	87
6	Conclusions and Future Work	88
A	Proof of Theorem 1	99

Acknowledgments

First, I would like to take this opportunity to thank my adviser, Professor Claire Tomlin, for her continued support and guidance. Throughout my PhD program, her insight, enthusiasm and dedication to her work never ceased to inspire me. In addition, I would like to thank Professor Anil Aswani for his advice and guidance, especially during the time that I was most frustrated and uncertain about my research direction. Furthermore, for their time and for their input, I thank Professor Murat Arcak for being on my thesis committee and I thank Professor Laurent El Ghaoui for being on my quals committee.

Despite the Hollywood image of a lone scientist, research is not a solitary process. Throughout my PhD program, I had the privilege to work with many superb collaborators. Whenever “we” appears in this thesis, I mean Young Hwan, Dariush, Datong, Vaggelis, Max, Frauke and Sumedh. In addition, there are many people that I collaborated with but our work is not included in this thesis: Jeff, Maja, Mo, Kene, Jaime, Casey, Gabriella, Anil, Mark and Soile. I learnt a great deal from working with everyone. I would also like to thank Domenico for facilitating the field experiments on SDH.

In addition, research is not the entirety of a PhD program. I would like to thank many caring, generous and fun people, who kept my spirit up during the past few years and gave me the courage and strength to finish my PhD: Can, Lucas, Sissi, TJ, Ming, Mo, Jeannette, Anil, Jerry and everyone else I missed. Finally, I would also like to thank my parents for supporting me in all my endeavors, from quitting my job to go to grad school to moving to a continent far away from home. I would also like to thank them for their help in taking care of Lucas, so that I can even find time to write up this thesis.

Chapter 1

Introduction

The electric power grid is a complex system subject to natural disasters and nuisance attacks, in addition to the rapid system dynamics and demand swings inherent in providing electric power across large areas. A balance between electricity generation and consumption at all times is one necessary condition for the normal operation of the power system, and large imbalances can cause power outages, leading to economic losses, physical damage, or even bodily harm. In addition, the power system is an example of a cyber-physical system (CPS), which consists of physical components such as actuators, sensors and controllers that communicate with each other over a network. Although communication networks are often protected by security measures, cyber attacks can still take place when a malicious attacker obtains unauthorized access. In a power system, cyber attacks not only compromise information, but can also cause physical damage. Therefore, it is desirable to protect the system against such cyber attacks. This thesis presents progress in the path towards finding solutions to these two problems and consists of two main parts. Chapters 2 and 3 describe first attempts at using commercial buildings to provide frequency regulation – a type of reserve used by electricity grid operators to balance electricity generation and demand. Chapters 4 and 5 focus on developing methods that estimate the true state of a power system when it is under cyber attack.


1.1 Frequency Regulation from Commercial Buildings

A balance of electricity generation and demand must be maintained at all times to achieve the reliable operation of the power system. Any mismatch between them is reflected through the grid frequency: if generation exactly meets demand, then the grid frequency is at its nominal value of 60 Hz (in the U.S.); if generation exceeds demand, then the frequency increases, and vice versa. Therefore, to maintain normal operation of the power grid, the grid operator uses reserves known as ancillary

services (AS) to correct any mismatch between electricity generation and demand. Amongst these reserves, frequency regulation is the highest quality AS over which the grid operator has almost real-time control and is active continuously during normal operation of the grid. Recent rapid increase in the penetration of renewable energy sources has aggravated the volatility and uncertainty of electricity generation, which led to a greater demand for frequency regulation reserves. Traditionally, these reserves have been provided by fast ramping power generators. However, an alternative is to explore the flexibility on the demand side, which may have less economic and environmental cost in the long run. More specifically, loads can provide regulation by increasing (decreasing) their electricity consumption when the grid frequency increases (decreases).

In particular, commercial buildings are a tremendous untapped resource for this application. First, they account for a large fraction of the total electricity consumption (more than 35% in the U.S., 39% of which is due to heating, ventilation and air conditioning (HVAC) systems [91]). Second, the building's large thermal capacity allows the power consumption of HVAC systems be partly shifted in time without compromising occupant comfort. Third, many commercial buildings are equipped with a variable frequency drive which can be controlled to vary the power consumption of supply fans of the HVAC system quickly and continuously [38]. This greatly simplifies tracking of the reference regulation signal, as opposed to resources with on-off control. Fourth, about one third of commercial buildings in the U.S. are equipped with a building automation system (BAS) [8] which facilitates the implementation of new controllers.

On the other hand, there are a number of challenges in using commercial buildings for frequency regulation. First, obtaining a building model that is amenable to control is not straightforward, because commercial buildings are often subject to large disturbances such as occupancy, that are difficult to capture. In addition, buildings are often not sufficiently excited, as they must satisfy strict regulatory requirements during regular operation, which limits the type and duration of excitation experiments that can be conducted. Second, about one third of commercial buildings in the U.S. are equipped with variable air volume (VAV) HVAC systems [38], which are typically complex with many control variables and interdependent control loops.

This thesis proposes procedures to develop both data-driven and physics-based models for the thermodynamic behavior of commercial buildings, and provides a quantitative comparison between them using both open- and closed-loop metrics. In addition, this thesis presents an experimental demonstration of the feasibility of using a VAV HVAC system for frequency regulation, where experiments are conducted in full accordance with Pennsylvania, New Jersey, Maryland's (PJM's) requirements. To the best of our knowledge, this is the first report where an occupied commercial building equipped with a VAV HVAC system  successfully provide frequency regulation.

1.2 Secure Estimation against Cyber Attacks

A key element in the development of smart power transmission systems over the past decade is the tremendous advancement of the synchrophasor technology. This is enabled by phasor measurement units (PMUs) which record and communicate Global Positioning System (GPS)-synchronized, high sampling rate dynamic power system data, and they are currently being installed at different points in the North American grid, especially under the smart grid initiatives of the U.S. Department of Energy. Significant efforts have been made in using PMU measurements for wide area control in a smart grid. In such a system, a wide area control system (WACS) communicates with PMUs over a communication network to achieve increased efficiency and reliability of the power system. However, communication networks are often vulnerable to cyber attacks. For example, [56] describes a multi-switch attack, in which different switches in a power network attacked at different times, can lead to stealthy and wide-scale cascading failures in the power system. Therefore, in order to protect the system against cyber attacks, the WACS must estimate the system's true states before using the received data for computing control signals. However, this is a challenging task as cyber attacks can be erratic and difficult to model.

Secure estimation problems study how to estimate the true system states when measurements are corrupted and/or control inputs are compromised by attackers. There has been tremendous amount of work in developing secure estimation algorithms, mostly for linear dynamical systems and/or by assuming the attack signal follows a certain distribution. However, the power system can only be approximated by a linear model under small perturbations. Under a severe disturbance, such as a single or multi-phase short-circuit or a generator loss, the linearized model does not remain valid [51], [95]. Therefore, the existing linear estimation techniques lack performance guarantees when the system undergoes large perturbations which are typical of highly loaded practical systems. To overcome the above drawbacks, this thesis focuses on sensor attacks and builds on previous results to first propose a secure state estimation algorithm for linear dynamical systems without any assumption on the sensor attacks or corruptions (i.e., corruptions can follow any particular model). The only assumption concerning the corrupted sensors is about the number of sensors that are corrupted due to attacks or failures. This thesis then extends the results to two classes of nonlinear systems and demonstrates through numerical simulations how the proposed estimation algorithm can be used to protect the nonlinear power system against cyber attacks.

Part 1

Commercial Buildings for Frequency Regulation

Chapter 2

Mathematical Modeling of Commercial Buildings

According to [77], commercial buildings account for more than 35% of the total electricity consumption in the U.S., with an upward trend. HVAC systems are a major source of this consumption [91]. Nevertheless, their power consumption can be flexibly scheduled without compromising occupant comfort, due to the thermal capacity of buildings. As a result, commercial buildings' HVAC systems have become the focal point of research, with the goal of utilizing this source of consumption flexibility. From the point of view of energy efficiency, researchers have studied optimization of building control in order to minimize power consumption [75, 85]. More recently, it has been proposed to engage buildings in supporting the supply quality of electricity and the grid stability, by participating in the regulation of electricity's frequency [4, 5, 55, 94].

All of the above research activities are based on a valid mathematical model describing the thermal behavior of buildings. Such a model should be identified using actual experimental data and capture realistic disturbances. In addition, it should describe the building's temperature evolution with suitable accuracy and spatial granularity, without being too computationally expensive. However, there are many challenges in identifying such a building model. First, buildings often have different types of spaces that are subject to very different uncertainties. Second, disturbances such as occupancy are difficult to predict *a priori*. Third, buildings are often not sufficiently excited, as they must satisfy strict regulatory requirements during regular operation, which limits the type and duration of excitation experiments that can be conducted.

In this chapter, we present two methods for the identification of a building model: a physics-based method and a data-driven method. We use both techniques to identify building models for the same testbed using experimental data collected from the building, and quantitatively compare the models using various metrics, including open-loop prediction accuracy and closed loop control strategies.

This chapter is an adaptation of the paper in [100].


2.1 Introduction

Traditionally, buildings have been modeled with high-dimensional physics-based models such as resistance-capacitance (RC) models [36, 61, 87, 88], TRNSYS [18] and EnergyPlus [97]. These models are motivated by the thermodynamics of the building and explicitly model the heat transfer between components of the buildings. The advantage of such models is their high granularity of temperature modeling, but a drawback is their high dimensionality which makes them computationally expensive. Although there has been extensive work on model reduction, this remains to be a non-trivial task. A large body of this work focuses on linear models, whereas physics-based models for commercial buildings with a VAV HVAC system are bilinear in nature. Furthermore, existing model reduction techniques often result in a loss of interpretability of states [16] and a significant increase in the model’s prediction error [31].

Motivated by these shortcomings, a new direction of research attempts to identify lower-dimensional, data-driven models, e.g. with Input-Output models [55] and semiparametric regression [2]. The purpose is to alleviate the computational complexity in expense for coarser and less accurate temperature predictions.

In this chapter, we propose both a physics-based method and a data-driven method to identify models of a multi-zone building, that is easy to implement with the building in regular operation, and captures internal gains such as occupancy, without the need of additional instruments like carbon dioxide sensors. Our procedures use excitation experiments that actively perturb the building and generate data that can be used for more accurate parameter identification.

More importantly, we perform a *quantitative comparison* of the data-driven and physics-based models in terms of open-loop prediction accuracy and closed-loop control strategies, based on the *same testbed* (the entire floor of an office building). We conclude that a low-dimensional data-driven model is suitable for building control applications, such as frequency regulation, due to its minor loss of prediction accuracy compared to high-dimensional physics-based models, but significant gain in computational ease. To the best of our knowledge, the extant body of literature has analyzed data-driven and physical models for the identification of temperature evolution in commercial buildings only in an isolated fashion (in particular not on the same testbed) [60], [85], [55], [42]. In addition, some of these models were identified for fictitious buildings with synthetic data [13, 32, 87], while others used experimental data collected under environments with little or no disturbance, e.g. without occupants [55]. Our work differs from these existing works in two aspects. First, we use experimental data to identify models for a multi-zone commercial building under regular operation, which is subject to significant disturbances such as occupancy. Second, although the existing literature mentions the differences between data-driven and physics-based models, the prevailing isolationist approach does not provide any quantitative comparison with respect to building control applications - a fact we

would like to alleviate by juxtaposing a data-driven with a physics-based model. 

2.2 Preliminaries

2.2.1 Building Description

We model the temperature evolution of the fourth floor of Sutardja Dai Hall (SDH), a seven-floor-building on the University of California, Berkeley campus. The fourth floor has a total area of 1300 square-meters and contains offices for research staff and open workspaces for students, and is divided into six zones for modeling purposes (Figure 2.1): Northwest (NW), Northeast (NE), West (W), Center (C), East (E), South (S).

The building is equipped with a VAV HVAC system that is common to 30% of all U.S. commercial buildings [90]. The system contains air handling units (AHU), inside which large supply fans drive air through heat exchangers, cooling it down to a desired supply air temperature (SAT). The cooled air is then distributed to VAV boxes located throughout the building. The flow rate and the final temperature of the supply air delivered to each room is then controlled by adjusting the damper position and the amount of reheating performed at the VAV boxes.

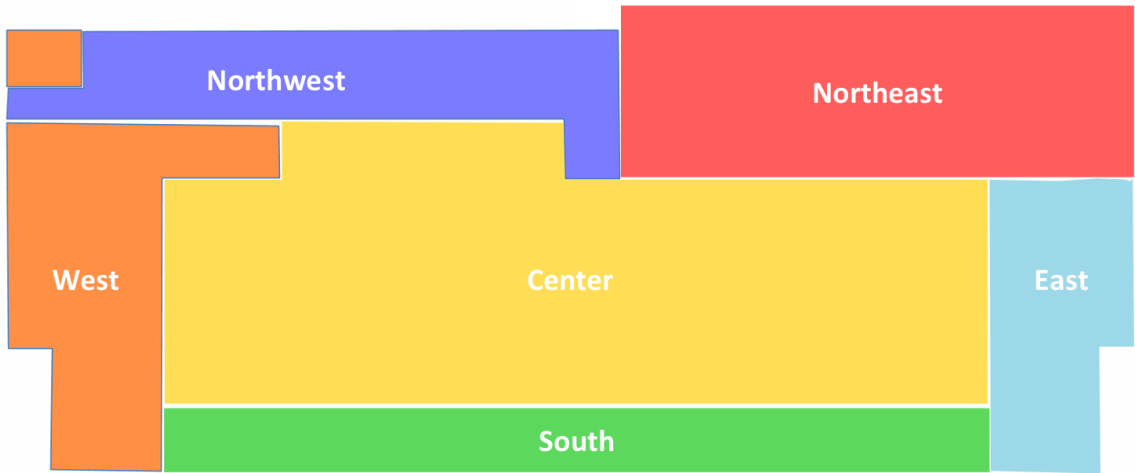


Figure 2.1: Zones for the fourth floor of Sutardja Dai Hall (SDH).

2.2.2 Collection of Experimental Data

We collected 51 weeks of one-minute resolution temperature data for the six zones along with the airflow rates of all 21 VAV boxes on the fourth floor, SAT

and the outside air temperature from the *simple Measurement and Actuation Profile* (sMAP). sMAP is a protocol that collects, stores and publishes time-series data from a wide variety of sensors [14,15]. The hourly global horizontal solar radiation data was obtained from a nearby weather station [70], from which the incidence solar radiation of the four geographic directions was calculated with the PV_LIB toolbox [86]. All collected data were down-sampled or interpolated, respectively, to 15 minute intervals. These 51 weeks of data span periods when the building was under normal operation as well as periods with excitation experiments.

To increase identifiability of the building model, forced response experiments were performed. These experiments were conducted during Saturdays to (a) minimize effects due to occupancy on our collected data, and thus facilitate subsequent parameter identification; (b) minimize impact to building operation and exploit larger comfort bounds on room temperatures during the weekends. Indeed, the comfort bounds were never violated during the forced experiments. Because of commercial buildings' large thermal inertia, each forced excitation must last sufficiently long before temperature changes are observable (we chose 2 hours for our excitation experiment). More specifically, starting at 8 a.m., the supply air's flow rate to one zone is set to its maximum value every 2 hours. During this 2 hour period, all of its adjacent zones' airflow rates are set to their minimum values, and each of the remaining zones' airflow rate is set to a random value. This procedure is repeated for each of the six zones.

2.2.3 Data Splitting

Next, we define the seasons “fall” (early September until mid December), “winter” (mid December until late January), and “spring” (late January until mid May) in order to account for different occupancy levels during the fall and spring semesters, and the winter break. After the weeks have been assigned to the seasons, a random portion of the data in each season (e.g. we chose 90%) was defined as the training data, and the remaining weeks to be removed prior to the analysis were declared as the test set, which were used to assess the accuracy of the building model fitted on the training data.

2.3 Data-Driven Model

We identify a difference equation for the temperature evolution with semiparametric regression, using 51 weeks of experimental data collected from the fourth floor of SDH. Semiparametric regression in buildings has been proposed by [2], where the authors used only one week of historic data to model the temperature evolution and used the HVAC's SAT as the single control input. We extend this approach by taking into account multiple weeks, which we separate into three seasons (fall, winter, spring) so as to characterize the different levels of the exogenous heating load for dif-

ferent temporal seasons. In addition, we model the room temperatures as a function of airflow rates from multiple VAVs to obtain a model which can be used for more sophisticated control strategies.

Next, we first introduce the semiparametric method using a simple lumped zone model of the fourth floor, and then identify a multi-zone model for the same floor which we will use to make quantitative comparisons with the physics-based model.

2.3.1 Lumped Zone

2.3.1.1 Model Setup

In order to facilitate analysis, the entire fourth floor of SDH is treated as a single zone, with the scalar temperature x corresponding to the average temperature on the entire floor and the input u as the sum of the inflow of all 21 VAV boxes. Then, the temperature evolution is assumed to have the following form:

$$x(k+1) = ax(k) + bu(k) + c^\top v(k) + q(k) + \epsilon(k), \quad (2.1)$$

where $v := [v_{Ta}, v_{Ts}, v_{solE}, v_{solN}, v_{solS}, v_{solW}]^\top$ is a vector of known disturbances that describes ambient air temperature, the HVAC system's supply air temperature, and solar radiation from each of the four geographical directions. In addition, q represents the internal gains due to occupancy and electric devices, and ϵ denotes independent and identically distributed zero mean noise with constant and finite variance which is conditionally independent of x , u , v , and q . Finally, $a, b \in \mathbb{R}$ and $c \in \mathbb{R}^6$ are unknown coefficients to be estimated using semiparametric regression [39, 81].

2.3.1.2 Smoothing of Time Series

The q term of Equation (2.1) is treated as a nonparametric term, so that (2.1) becomes a partially linear model. By taking conditional expectations on both sides of (2.1), we obtain

$$\hat{x}(k+1) = a\hat{x}(k) + b\hat{u}(k) + c^\top \hat{v}(k) + \mathbb{E}[q(k)|k] + \mathbb{E}[\epsilon(k)|k], \quad (2.2)$$

where the conditional expectations $\hat{x}(\cdot) = \mathbb{E}[x(\cdot)|\cdot]$, $\hat{u}(\cdot) = \mathbb{E}[u(\cdot)|\cdot]$, and $\hat{v}(\cdot) = \mathbb{E}[v(\cdot)|\cdot]$ are used. Noting that $\mathbb{E}[\epsilon(\cdot)|\cdot] = 0$ and assuming $\mathbb{E}[q(\cdot)|\cdot] = q(\cdot)$, subtracting (2.2) from (2.1) gives

$$\begin{aligned} x(k+1) - \hat{x}(k+1) &= a(x(k) - \hat{x}(k)) + b(u(k) - \hat{u}(k)) \\ &\quad + c^\top (v(k) - \hat{v}(k)) + \epsilon(k). \end{aligned} \quad (2.3)$$

The unknown internal gains term has been eliminated, and thus the coefficients a, b, c in (2.3) can be estimated with any regression method.

The conditional expectations $\hat{x}(\cdot)$, $\hat{u}(\cdot)$ and $\hat{v}(\cdot)$ are obtained by smoothing the respective time series [2]. We made use of locally weighted linear regression with a tricube weight function, where we use k -fold cross-validation to determine the bandwidth for regression. The error measure used for in-sample estimates is the Root Mean Squared (RMS) error between the measured temperatures $\bar{x}(k)$ and the model's predicted temperatures $x(k)$ over a time horizon of N steps (e.g. we chose a 24 hour time horizon, $N = 96$):

$$\text{RMS error} = \left(\frac{1}{N} \sum_{k=1}^N [\bar{x}(k) - x(k)]^2 \right)^{1/2}. \quad (2.4)$$

2.3.1.3 Bayesian Constrained Least Squares

A main challenge in identifying the model is that commercial buildings are often insufficiently excited. Take SDH for example, whose room temperatures under regular operation only vary within a range of 2°C and inflow of the VAV boxes hardly varies at all. To overcome this, data collected during forced response experiments described in Section 2.2.2 was used in training the model. To further compensate for the lack of excitation, a Bayesian regression method is used, which allows our prior knowledge of the building physics to be incorporated in the identification of coefficients. More specifically, Gaussian prior distributions are used for the coefficients a and b , i.e., $a \sim \mathcal{N}(\mu_a, \Sigma_a)$ and $b \sim \mathcal{N}(\mu_b, \Sigma_b)$, where $\mathcal{N}(\mu, \Sigma)$ denotes a jointly Gaussian distribution with mean μ and covariance matrix Σ . In addition, a , b and c are constrained to be identical for the different seasons, since they model the underlying physics of the building which are assumed to be invariant throughout the year.

Let $\mathcal{T} = \{1, 2, \dots, N\}$ denote the N weeks of training data (e.g. we chose $N = 45$), and define $\mathcal{F} = \{i \in \mathcal{T} \text{ such that } i \text{ is a week in fall}\}$ as the set of training weeks from the fall season. Similarly, define the sets of training weeks from the winter and spring as \mathcal{W} and \mathcal{S} . The coefficient identification problem is then formulated as follows:

$$\begin{aligned} (\hat{a}, \hat{b}, \hat{c}) &= \arg \min_{a, b, c} (J_{\mathcal{F}} + J_{\mathcal{W}} + J_{\mathcal{S}}) + \|\Sigma_a^{-1/2}(a - \mu_a)\|^2 + \|\Sigma_b^{-1/2}(b - \mu_b)\|^2 \\ \text{s.t. } J_{\mathcal{X}} &= \sum_{i \in \mathcal{X}} \|x_i(k+1) - \hat{x}_i(k+1) - a(x_i(k) - \hat{x}_i(k)) \\ &\quad - b(u_i(k) - \hat{u}_i(k)) - c^\top(v_i(k) - \hat{v}_i(k))\|^2 \quad \text{for } \mathcal{X} \in \{\mathcal{F}, \mathcal{W}, \mathcal{S}\}, \\ 0 &< a < 1, \quad b \leq 0, \quad c \geq 0. \end{aligned} \quad (2.5)$$

In other words, $J_{\mathcal{F}}$, $J_{\mathcal{W}}$ and $J_{\mathcal{S}}$ represent the sum of squared errors between experimentally measured temperatures and model's predicted temperatures for fall, winter and spring, respectively. The sign constraints on the parameters b and c translate into the fact that the temperature to be estimated positively correlates with all components in v and negatively correlates with the VAV airflow. The range of a is a consequence of Newton's Law of Cooling.

To find the effect of the VAV inflow on the 15-minute temperature evolution, we computed the 15-minute incremental reductions in temperature Δx recorded during the excitation experiments. It is assumed that the large inflow u dominates all other effects such that we can assume

$$\Delta x = x(k+1) - x(k) = b \cdot u(k) \quad (2.6)$$

for all k during the excitation period. The estimated prior μ_b can then be isolated from (2.6). The prior μ_a was set as the optimal \hat{a} identified by (2.5) without the prior terms. The covariance matrices Σ_a and Σ_b were chosen subjectively.

2.3.1.4 Estimation of Internal Gains

With the estimated coefficients $\hat{a}, \hat{b}, \hat{c}$ in hand, the internal gains q can be estimated by manipulating (2.2):

$$\hat{q}(k) = \hat{x}(k+1) - \left(\hat{a}\hat{x}(k) + \hat{b}\hat{u}(k) + \hat{c}^\top \hat{v}(k) \right). \quad (2.7)$$

A distinct function of internal gains is estimated for each season. In other words, (2.7) is used to estimate an instance of the internal gains function for each week i in the training set \mathcal{T} . The internal gains function for each season, $\hat{q}_{\mathcal{X}}$ is then defined as the average of estimated weekly gains for all weeks $i \in \mathcal{X}$ and $\mathcal{X} \in \{\mathcal{F}, \mathcal{W}, \mathcal{S}\}$.

2.3.1.5 Results

The estimated internal gains for each season are shown in Figure 2.2. Observe

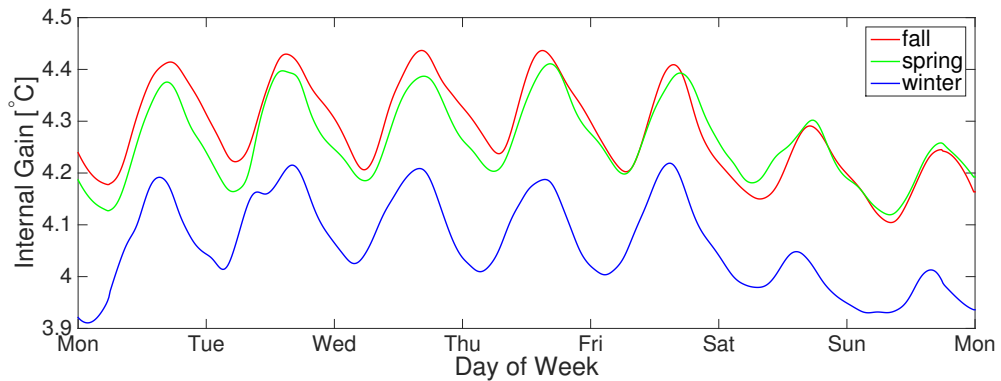


Figure 2.2: Estimated internal gain q from the data-driven model by season, lumped case.

that, for all three seasons, the internal gains exhibit a daily trend with local peaks around the late afternoon and local minima at night. Moreover, the amplitudes of

the internal gains are considerably smaller during the weekends, suggesting a lighter occupancy. It can further be seen that the magnitude of the internal gains is smallest for the winter season, which is in accordance with our intuition since most building occupants are absent during that period.

Lastly, since the Bayesian Constrained Least Squares algorithm (2.5) has identified a set of parameter estimates $\hat{a}, \hat{b}, \hat{c}$ valid for all three seasons to account for the time-invariant physics of the building, the temperature predictions are of the same nature for all three seasons. We thus conclude that the inherent differences between the seasonal temperature data are captured by the internal gains and can be compared between the seasons on a relative level.

The identified models for the different seasons found with (2.5) are

$$\begin{aligned}
 x(k+1) &= 0.80 \cdot x(k) - 0.18 \cdot u(k) \\
 &\quad + [0.0019, 0.028, \mathbf{0}] v(k) + q_{\mathcal{X}}(k) \\
 &= 0.80 \cdot x(k) - 0.18 \cdot u(k) \\
 &\quad + 0.0019 \cdot v_{\text{Ta}}(k) + 0.028 \cdot v_{\text{Ts}}(k) + q_{\mathcal{X}}(k) \\
 &\text{for } \mathcal{X} \in \{\mathcal{F}, \mathcal{W}, \mathcal{S}\}.
 \end{aligned} \tag{2.8}$$

The estimated coefficients of c corresponding to the solar radiation disturbances are very small ($< 10^{-6}$) compared to the other estimated coefficients. Since the temperatures are of the order 10°C , air inflow around 1 kg/s and solar radiation about 100 W/m^2 , the effect of solar radiation on the room temperature is orders of magnitude less than that of other factors and hence can be neglected.

The average RMS prediction errors are 0.22°C , 0.17°C and 0.23°C for fall, winter and spring, respectively, showing that our model predicts the temperature reasonably well.

2.3.2 Individual Zones

2.3.2.1 Model Setup

Rather than approximating the entire fourth floor of SDH as a single zone, in this section, we identify a multivariate model that describes the thermodynamic behavior of each of the six individual zones:

$$\begin{aligned}
 x(k+1) &= Ax(k) + Bu(k) + Cv(k) + q_{\mathcal{X}}(k) \\
 &\text{for } \mathcal{X} \in \{\mathcal{F}, \mathcal{W}, \mathcal{S}\},
 \end{aligned} \tag{2.9}$$

where $x, q_{\mathcal{X}} \in \mathbb{R}^6$, and the control input $u \in \mathbb{R}^6$ represent the temperatures, the internal gains of each zone, and the total air flow to each zone, respectively. In the lumped case, it was observed that solar radiation only had a negligible effect on the building's thermodynamics compared to the input and other disturbances, and thus we omit the solar radiation in the subsequent analysis: $v := [v_{\text{Ta}}, v_{\text{Ts}}]^\top \in \mathbb{R}^2$.

Inspired by Newton’s Law of Cooling, only adjacent zones influence each other’s temperature, which defines the sparsity pattern of the coefficient matrices that are to be estimated. Hence

$$A_{ij} = \begin{cases} \neq 0, & \text{if } i = j \text{ or } (i, j) \text{ adjacent} \\ 0, & \text{otherwise.} \end{cases} \quad (2.10)$$

The diagonal elements of A denote autoregressive terms for zone temperatures, whereas non-diagonal elements describe the heat exchange between adjacent rooms. The matrix B is diagonal by definition of u . The sparsity pattern of C is found by physical adjacency of a respective zone to an exterior wall of a given geographic direction.

2.3.2.2 Model Identification

The procedure for the estimation of the parameter matrices \hat{A} , \hat{B} , \hat{C} , and the internal gains follows (2.5), but with a modified choice of the (now matrix-valued) priors μ_a and μ_b : μ_b and the diagonal entries of μ_a are obtained by scaling the corresponding priors from the lumped zone case in order to account for the thermal masses of the individual zones, which are smaller than in the lumped case. The off-diagonal elements of μ_a , which represent the heat transfer between adjacent zones, were set to a value close to zero, according to our calculations with the heat transfer equation $\dot{q} = U \cdot A \cdot \Delta x$ and [49].

2.3.2.3 Results

Figure 2.3 shows the estimated internal gains for the three seasons fall, winter, and spring for the six single zones, computed with the smoothed time series (2.7). It can be seen that the different zones exhibit different magnitudes of internal gains, with average values of the internal gains ranging between 1.0°C and 3.6°C for different zones and seasons. Similar to the lumped zone case (Figure 2.2), daily peaks of the internal gains profiles can be recognized, with a slight decrease in magnitude on weekend days. The average prediction RMS error by zone and season are reported in Table 2.2.

2.4 Physics-Based Model

In this section, we identify a difference equation for the temperature evolution using the Resistance-Capacitance (RC) modeling method, via the Building Resistance-Capacitance Modeling (BRCM) MATLAB toolbox [87]. To derive an RC building model, we first decompose the building into building elements (BE), such as the bulk volume of air in each room, walls, floors and ceilings. An electric analogy can then be used to obtain an equivalent electrical circuit whose resistances and capacitances

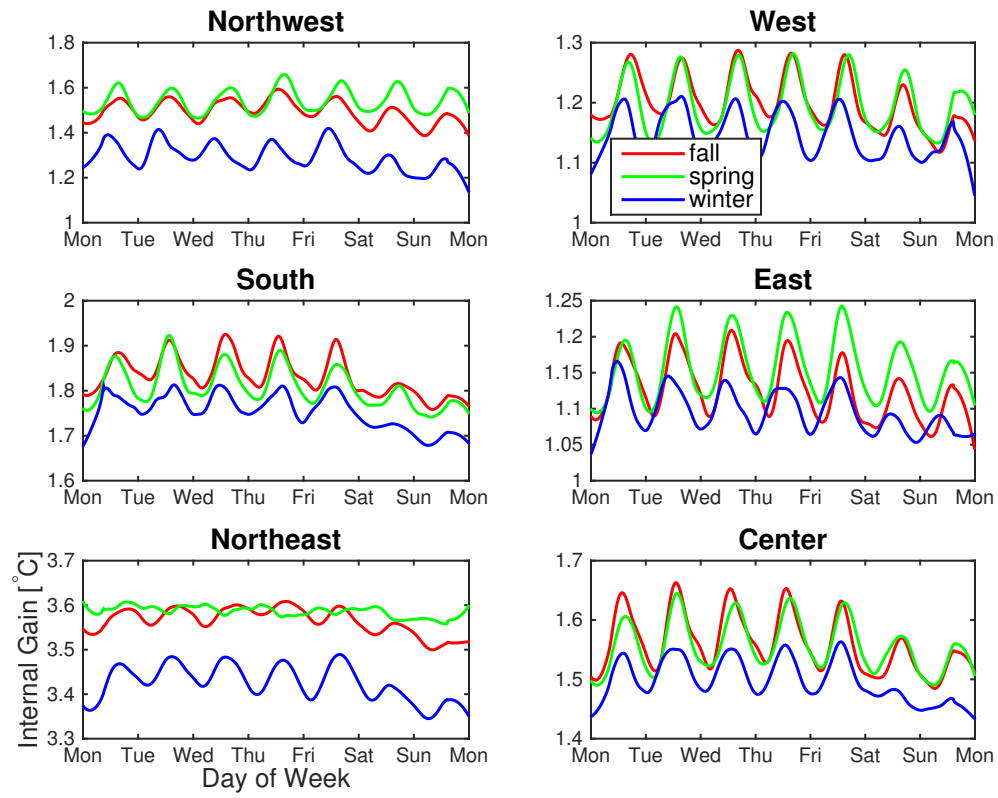


Figure 2.3: Estimated internal gain q from the data-driven model by zone and season, individual case.

represent thermal resistances and thermal capacitances of the BEs, and voltages and currents represent temperatures of BEs and heat transfers between those. The thermal dynamic model can be obtained by applying Ohm's law and Kirchhoff's circuit laws on the equivalent circuit. The resulting building model is bilinear in nature, due to the physics of the HVAC system.

2.4.1 Model Setup

The physics-based building model has the following form [42]:

$$x(k+1) = Ax(k) + B_v v(k) + B_q q(k) \quad (2.11a)$$

$$+ \sum_{i=1}^{21} (B_{xu_i} x(k) + B_{vu_i} v(k)) u_i(k) \\ y(k) = Cx(k), \quad (2.11b)$$

where the state vector $x \in \mathbb{R}^{289}$ represents temperatures of all building elements on the fourth floor and $y \in \mathbb{R}^6$ represents the average temperatures of the six zones shown in Figure 2.1. $u \in \mathbb{R}^{21}$ denotes the airflow rate from the 21 VAV boxes, and $v := [v_{Ta}, v_{Ts}]^\top$ is the disturbance vector, which captures known disturbances from ambient air temperature and the SAT. Note that from Section 2.3, heat gains due to solar radiation are orders of magnitude less than those caused by other disturbances and inputs, hence are not included here. Finally, $q(k) : \mathbb{N} \rightarrow \mathbb{R}^6$ captures internal gains in each of the six zones on the fourth floor. For week m from the training set \mathcal{T} :

$$q(k) = q_0 + \begin{cases} q_{\mathcal{F}}(k), & \text{if } m \in \mathcal{F}, \\ q_{\mathcal{W}}(k), & \text{if } m \in \mathcal{W}, \\ q_{\mathcal{S}}(k), & \text{if } m \in \mathcal{S}, \end{cases} \quad (2.12)$$

where q_0 is an unknown constant vector representing background heat gains due to idle appliances such as computers and printers. Functions $q_{\mathcal{F}}(\cdot)$, $q_{\mathcal{W}}(\cdot)$ and $q_{\mathcal{S}}(\cdot)$ are unknown nonparametric functions that capture the time-varying heat gain due to occupancy and equipments in fall, winter and spring, respectively. The system matrices A , B_v , B_q , B_{xu_i} and B_{vu_i} are functions of tuning parameters: the window heat transmission coefficient (U_{win}), the convection coefficients of the interior wall (γ_{IW}), the exterior wall (γ_{EW}), the floor (γ_{floor}), and the ceiling (γ_{ceil}). Define $\gamma := [U_{\text{win}}, \gamma_{\text{IW}}, \gamma_{\text{EW}}, \gamma_{\text{floor}}, \gamma_{\text{ceil}}, q_0^\top]^\top \in \mathbb{R}^{11}$, then to identify the physics-based model, we need to estimate the parameter vector γ as well as the functions $q_{\mathcal{X}}(\cdot)$, $\mathcal{X} \in \{\mathcal{F}, \mathcal{W}, \mathcal{S}\}$. Next, we describe our approach for identifying this model.

2.4.2 Model Identification

For a fair comparison, the same data used to train and test the data-driven model is used to train and validate the physics-based model. The model identification

process is performed in two steps. First, the subset of the training data collected during weekends is used to estimate the parameters, γ . Second, the nonparametric functions $q_{\mathcal{X}}(\cdot)$ are estimated from the complete training dataset.

2.4.2.1 Parameter Estimation

For parameter estimation purposes, we first set $q_{\mathcal{X}}(\cdot) = 0$ during the weekend days, and evaluate them at a later point (Equations (2.17)). With $q_{\mathcal{X}}(\cdot) = 0$, (2.11) reduces to a purely parametric model:

$$\begin{aligned} x(k+1) &= Ax(k) + B_v v(k) + B_q q_0 \\ &\quad + \sum_{i=1}^{21} (B_{xu_i} x(k) + B_{vu_i} v(k)) u_i(k), \\ y(k) &= Cx(k). \end{aligned} \quad (2.13)$$

The optimal model parameters are estimated by solving the following optimization problem:

$$\begin{aligned} \hat{\gamma} &= \arg \min_{\gamma > 0} \sum_{m \in \mathcal{T}} \sum_k \|y_m(k, \gamma) - \bar{y}_m(k)\|^2 \\ \text{s.t. } &y_m(k, \gamma) \text{ and } x_m(k, \gamma) \text{ satisfy (2.13) with} \\ &x_m(0) = x_{\text{KF},m}(0), \quad u_m(k) = \bar{u}_m(k), \quad v_m(k) = \bar{v}_m(k) \quad \forall k, \end{aligned} \quad (2.14)$$

where \bar{u} , \bar{v} and \bar{y} denote the measured inputs, disturbances, and zone temperatures, respectively. In other words, we choose γ such that, when the model is simulated with this set of parameter values and the measured inputs and disturbances, the sum of squared errors between the measured zone temperatures and the simulated temperatures is minimized. The initial state $x_m(0)$ is required to simulate the model, however, not all states are measurable (the wall temperature for example is not), thus we estimate the initial states using a Kalman Filter $x_{\text{KF},m}(0)$, and set $x_m(0) = x_{\text{KF},m}(0)$. Furthermore, to compensate for the lack of sufficient excitation of the building, initial guesses for γ that are physically plausible are chosen. The optimal parameter values are reported in Table 2.1.

2.4.2.2 Estimation of $q(\cdot)$ for Each Season

Let $q_m(\cdot)$, $m \in \mathcal{T}$ be an instance of the internal gains function $q(\cdot)$ estimated for week m in the training set. The optimal estimate for a given season, say fall, is then defined as the the average of all estimates for that season:

$$\hat{q}_{\mathcal{F}}(k) = \frac{\sum_{m \in \mathcal{F}} q_m(k)}{\|\mathcal{F}\|} \quad \text{for all } k, \quad (2.15)$$

where $\|\mathcal{F}\|$ represents the cardinality of set \mathcal{F} .

To estimate $q_m(\cdot)$ for a given week m , let $\tilde{x}(k)$ and $\tilde{y}(k)$ denote the predicted states and zone temperatures at time k , with $q_m(k-1) = 0$, i.e.,

$$\begin{aligned}\tilde{x}(k) &= Ax(k-1) + B_v v(k-1) + B_q q_0 \\ &\quad + \sum_{i=1}^{21} (B_{xu_i} x(k-1) + B_{vu_i} v(k-1)) \\ &\quad \cdot u_i(k-1), \\ \tilde{y} &= C\tilde{x}(k).\end{aligned}\tag{2.16}$$

By noting $x(k) = \tilde{x}(k) + B_q q_m(k-1)$, $q_m(k-1)$ can be estimated by solving the following set of linear equations using Ordinary Least Squares:

$$(CB_q) \cdot q_m(k-1) = \bar{y}(k) - \tilde{y}(k),\tag{2.17}$$

where $\bar{y}(k)$ is the measured zone temperatures at time k .

2.4.3 Results

The identified model is tested on holdout test weeks from different seasons. The average daily prediction RMS errors by zone and season are reported in Table 2.2. Figure 2.4 shows the estimated increase in zone temperatures due to internal gains for fall, winter and spring. Similar average internal gains are observed for all zones and seasons. The zones that correspond to open workspaces and conference rooms (West, South, East and Center) show discernible daily peaks in their internal gains profiles with a slight decrease during weekends. Furthermore, there is little variation in the internal gains profiles across different seasons.

2.5 Quantitative Comparison of both Models

2.5.1 Prediction Accuracy

The high-dimensional physics-based model (Model B) is found to have a higher prediction accuracy compared to the low-dimensional data-driven model for the individual zones (Model A) presented in Section 2.3.2: According to Table 2.2, the mean RMS error for Model B across zones is 0.11°C lower than for Model A. This is also illustrated in Figure 2.5, which shows 7-day open-loop predictions of the temperature of a randomly selected holdout test week in the spring period, simulated with both models initialized with the measured temperature. The increase in RMS error from Model B to Model A is notably larger in the zones East (0.38) and Center (0.15), compared to the other zones (0.11, -0.03 , 0.07, and 0.08).

This provides new insight into the existing knowledge as we provide a quantitative comparison between the low-dimensional data-driven model and the high-dimensional physics-based model's prediction accuracy for the same multi-zone com-

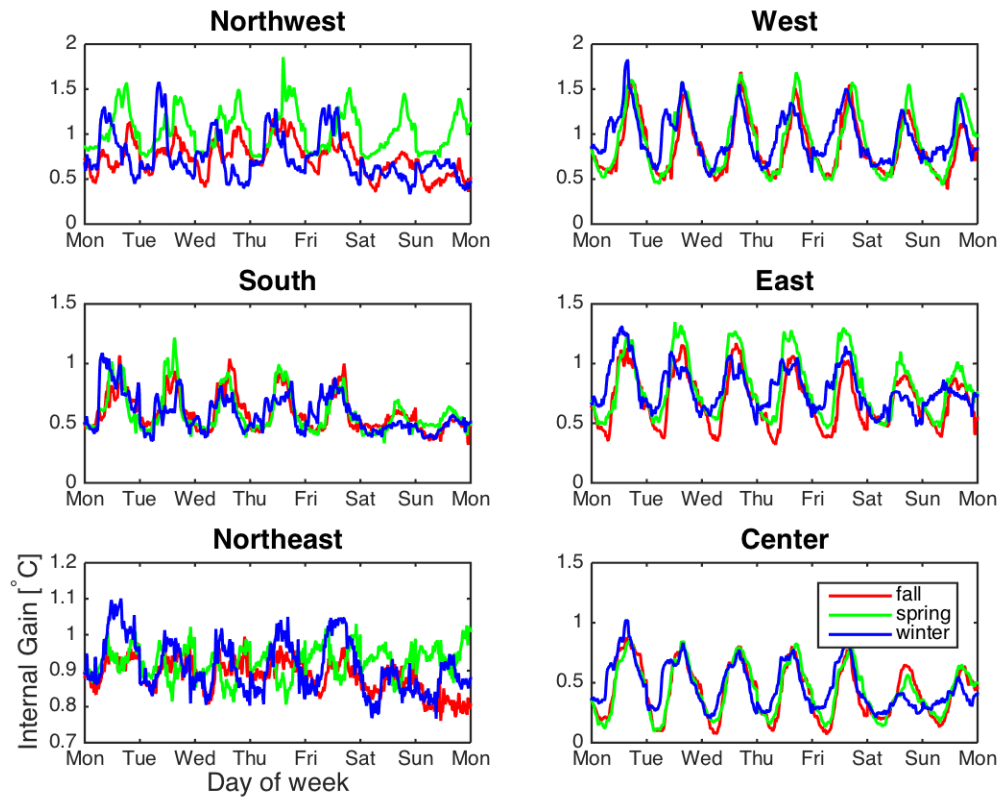


Figure 2.4: Estimated internal gain q from the physics-based model by zone and season.

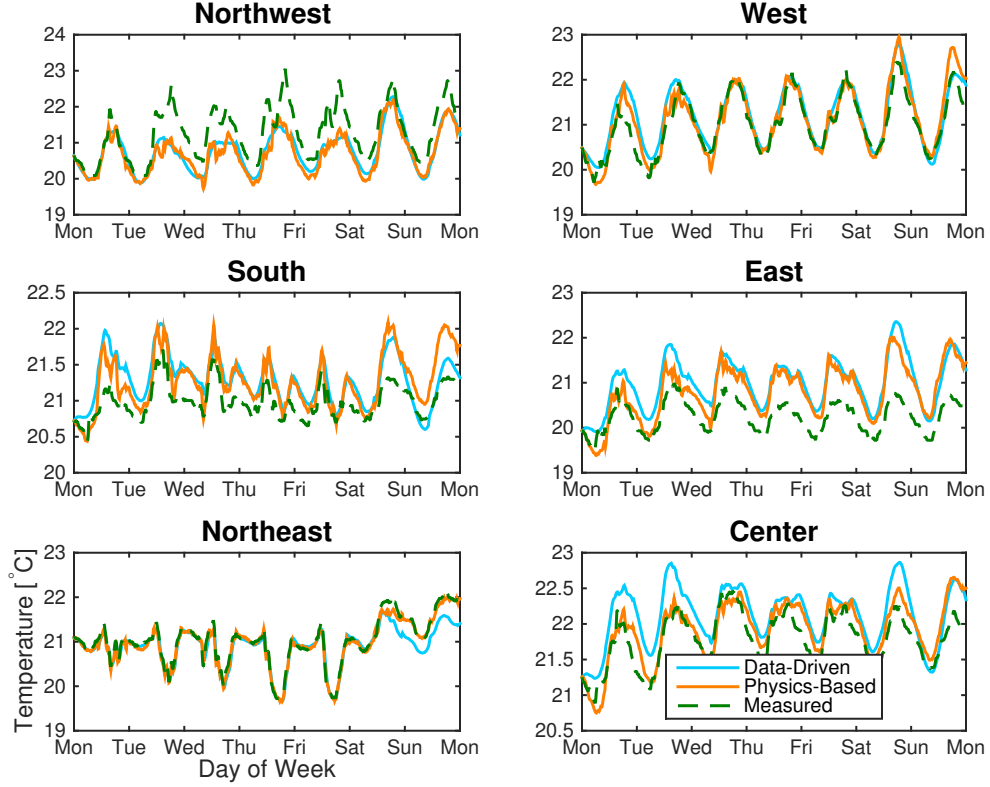


Figure 2.5: Simulated temperatures from the data-driven model (blue), physics-based model (orange) and actual temperatures (green).

mercial building, which is in regular operation. The existing literature merely mentions that data-driven models are likely to have lower prediction accuracies than physics-based ones and, to the best of our knowledge, a quantitative comparison at this level is non-existent, as previous building models were developed for different testbeds, fictitious buildings or from simulated data.

Next, we explore the extent to which this slightly lower prediction accuracy of Model A affects its resulting controller’s closed-loop performance in a building energy efficiency example.

2.5.2 Energy Efficient Control

In this section, we compare the performance of Model A and Model B for the purpose of energy efficiency. We formulate a model predictive control (MPC) problem to find the optimal control strategy that minimizes the cost of HVAC operation over the same week used in Figure 2.5, while guaranteeing the temperature to stay within a comfort zone $[T_{\min}, T_{\max}]$, which we chose as $[20^{\circ}\text{C}, 22^{\circ}\text{C}]$ [35], and confining the

control input to the physical limits of the HVAC system $[u_{\min}, u_{\max}]$. This problem is formulated as follows:

$$\begin{aligned}
& \min_{u, \varepsilon} \sum_{k=1}^N u(k)^2 + \rho \|\varepsilon\|_2 \\
& \text{s.t. } x(0) = \bar{x}(0) \\
& x(k+1) = \begin{cases} (2.9), & \text{Model A} \\ (2.11a), & \text{Model B} \end{cases} \\
& u_{\min} - \varepsilon \leq u(k) \leq u_{\max} + \varepsilon \quad \forall k \in [0, N-1] \\
& \begin{cases} T_{\min} \leq x(k) \leq T_{\max}, & \text{Model A} \\ T_{\min} \leq Cx(k) \leq T_{\max}, & \text{Model B (2.11b)} \end{cases} \quad \forall k \in [1, N]
\end{aligned} \tag{2.18}$$

The temperature is initialized with the measured temperature $\bar{x}(0)$ at the beginning of the week-long simulation. We use soft constraints on the control input with a penalty parameter ρ to ensure the feasibility of the problem. The penalty represents the cost of increasing the airflow beyond the operating limits (temporary shutdown or overuse, both of which are harmful to the system). To find the optimal control strategy, we make use of receding horizon control with a prediction horizon of three 15-minute time steps.

Figure 2.6 shows the temperature trajectory computed by the energy efficient controller (2.18) computed with both models A and B, together with the measured temperature as a reference. It can be seen that both control schemes are capable of maintaining the temperature within $[20^\circ\text{C}, 22^\circ\text{C}]$, with a control strategy that is of comparable cost (1,006 and 1,731 for Model A and Model B, respectively, where $\rho = 100$), shown in Figure 2.7. An interesting observation is that the largest difference in the control strategies is detected in zones East and Center, which show a larger increase in RMS error from Model B to Model A. Furthermore, it is interesting to observe that variations in the control input do not impact the periodicity of the temperature qualitatively, which can be explained by the regularity of the identified internal gains.

These findings suggest that both models perform equally well in designing an energy efficient control strategy. However, computing this strategy for Model A was cheap (< 5 minutes) compared to Model B (≈ 20 hours) on a 2 GHz Intel Core i7, 16 GB 1600 MHz DDR3 machine. Further, we note that in real-world applications, the MPC would use state feedback to initialize the temperature with sensor measurements at every time step, whereas in our simulation, it operates in an “open loop” fashion and hence propagates the estimation error with time. This will reduce the difference in the prediction quality by both controllers even further, since the RMS error is now to be evaluated on a much shorter prediction horizon, thereby further corroborating the finding of almost identical control schemes.

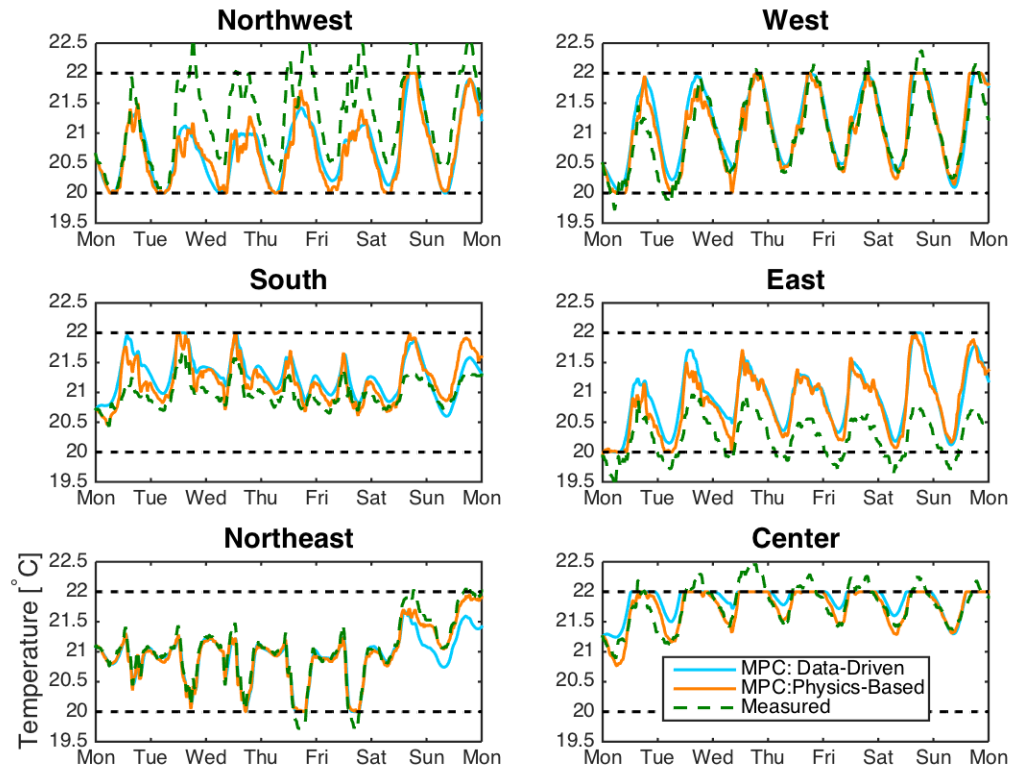


Figure 2.6: Optimal temperature for MPC with data-driven model (blue), MPC with physics-based model (orange) and actual temperature (green).

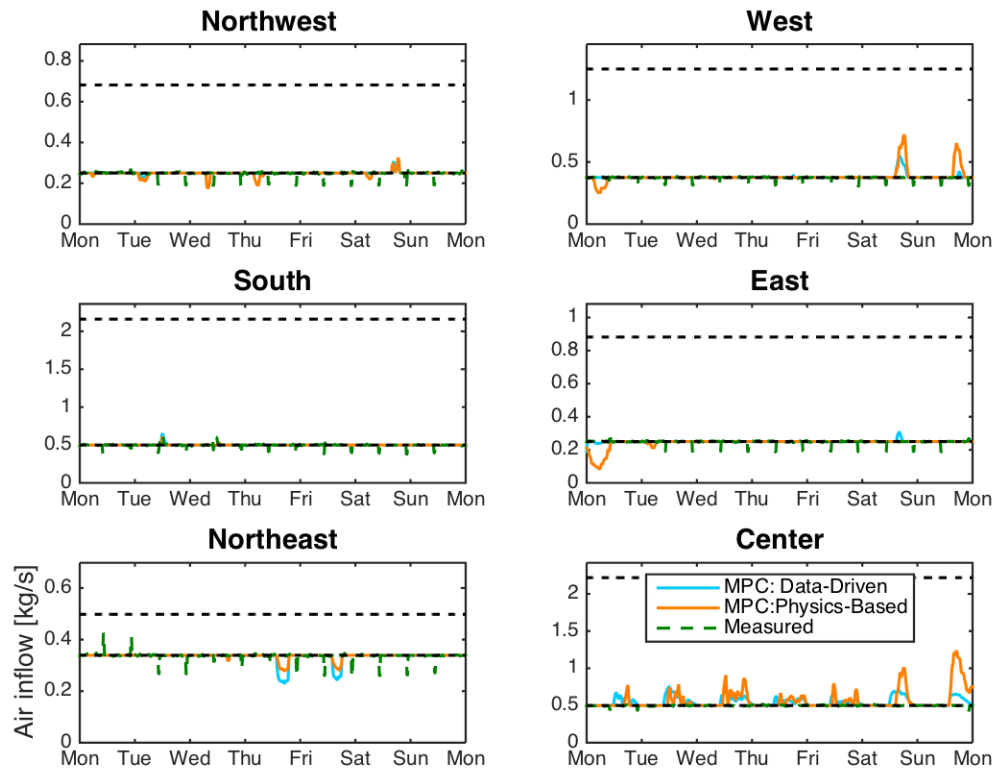


Figure 2.7: Optimal control strategy for MPC with data-driven model (blue), MPC with physics-based model (orange) and actual input (green).

Observing that Model A only suffers a negligible loss of accuracy compared to Model B for an open loop optimal control scheme, our findings suggest the applicability of Model A to other applications with temperature-critical zones in which even more precise temperature estimates are needed, e.g. long-term planning of reserve provision for frequency regulation.

2.6 Conclusion

We identified two state-space models for the thermal behavior of the same multi-zone commercial building using experimental data collected during regular building operation. One of the models is a low-dimensional data-driven model identified using semiparametric regression, the other one is a high-dimensional physics-based resistance-capacitance model. Both models capture the effect of disturbances such as occupancy and electrical appliances that commercial buildings are subjected to, without installation of any additional hardware such as occupancy sensors.

The identification of both models on the *same building* enabled us to quantitatively compare the performance of these types of models when applied to a real building, which has not been done before. Our results showed that the RMS error of the open-loop temperature prediction of the physics-based model across different thermal zones and temporal seasons is 0.11°C lower than in the data-driven model, a 25% reduction. However, simulating energy efficient MPC schemes under both models suggested both models perform equally well in terms of cost function minimization and constraint satisfaction despite the significantly higher complexity of the physics-based model.

It is widely known in this field that low-dimensional data-driven models have lower prediction accuracy than high-dimensional physics-based models, and thus have been only proposed for control of less temperature-critical buildings or zones. However, our work investigated an identification method for data-driven models for multi-zone commercial buildings in regular operation and demonstrated that the lower open-loop prediction accuracy of such data-driven models is insignificant in closed-loop control schemes compared to a high-dimensional physics-based model. Based on these findings, we suggest that such data-driven models may be suitable for applications that were previously considered inappropriate, e.g. frequency regulation.

Parameter	Description	Value [Unit]
γ_{EW}	Exterior Wall Convection Coeff.	50.0 [W/(m ² K)]
γ_{IW}	Interior Wall Convection Coeff.	12.8 [W/(m ² K)]
γ_{floor}	Floor Convection Coeff.	50.0 [W/(m ² K)]
γ_{ceil}	Ceiling Convection Coeff.	5.0 [W/(m ² K)]
U_{win}	Window Heat Transmission Coeff.	5.5 [W/(m ² K)]
$q_{0,NW}$	Background Heat Gain in Zone NW	25.6 [W/m ²]
$q_{0,W}$	Background Heat Gain in Zone W	20.6 [W/m ²]
$q_{0,S}$	Background Heat Gain in Zone S	15.9 [W/m ²]
$q_{0,E}$	Background Heat Gain in Zone E	16.4 [W/m ²]
$q_{0,NE}$	Background Heat Gain in Zone NE	19.9 [W/m ²]
$q_{0,C}$	Background Heat Gain in Zone C	6.8 [W/m ²]

Table 2.1: Optimal parameter values of physics-based model.

Data-Driven Model							
Season	NW	W	S	E	NE	C	Mean
Fall	0.98	0.61	0.28	0.42	0.28	0.36	0.488
Winter	1.41	0.34	0.29	0.26	0.25	0.21	0.460
Spring	0.56	0.25	0.31	0.71	0.17	0.34	0.390
Physics-Based Model							
Season	NW	W	S	E	NE	C	Mean
Fall	0.61	0.46	0.39	0.39	0.20	0.32	0.396
Winter	0.55	0.39	0.34	0.32	0.18	0.24	0.338
Spring	0.45	0.28	0.24	0.33	0.09	0.19	0.263

Table 2.2: RMS error by zone and season for data-driven and physics-based models.

Chapter 3

Experimental Demonstration of Frequency Regulation from Commercial Buildings

A balance of electricity generation and consumption at all times is one of the necessary requirements for the normal operation of a power system. Reserves known as ancillary services (AS) are used to correct any mismatch between generation and consumption. Amongst these reserves, frequency regulation is the highest quality AS over which the grid operator has almost real-time control and is active continuously during normal operation of the grid, to maintain the grid frequency at its nominal value (60 Hz in the U.S.). The recent rapid increase in the penetration of renewable energy sources has aggravated the volatility and uncertainty of electricity supply, which leads to a greater demand for frequency regulation reserves. These reserves have been traditionally provided by conventional fast-ramping power generators. More recently, loads on the demand side have also been considered for this application, in particular, commercial buildings. The feasibility of this proposal has been investigated by numerous researchers through simulations and theoretical studies. On the experimental side, there has only been a limited number of field tests, most of which were conducted in controlled laboratory environments with minimal uncertainties.

In this chapter, we demonstrate experimentally that commercial buildings equipped with HVAC systems can provide frequency regulation. First, we propose a control scheme that adjusts the electricity consumption of the supply fans of the HVAC system to track the frequency regulation signal. Then, we demonstrate the performance of our proposed control scheme through numerous field experiments conducted in an occupied commercial building, and in accordance with PJM's regulation market rules.

3.1 Introduction

Commercial buildings are a tremendous untapped resource for frequency regulation provision for various reasons. First, they account for a large fraction of the total electricity consumption (up to 40% in the U.S., up to 35% of which is due to HVAC systems [91]). Second, the electricity consumption of HVAC systems can be flexibly scheduled without compromising occupant comfort thanks to its large thermal inertia. Third, many commercial buildings are equipped with a variable frequency drive which can be controlled to vary the power consumption of supply fans of the HVAC system quickly and continuously [38]. This greatly simplifies tracking of the reference regulation signal, as opposed to resources with on-off control. Fourth, about one third of commercial buildings in the U.S. are equipped with a BAS [8] which facilitates the implementation of new controllers. On the other hand, commercial buildings are often subject to large disturbances and uncertainties, such as occupancy, which are difficult to capture and predict. In addition, about one third of commercial buildings in the U.S. are equipped with VAV HVAC systems [38], which are typically complex with many control variables and interdependent control loops.



Consequently, most of the existing literature on using commercial buildings for frequency regulation is based on simulation and theory. *Zhao et al.* demonstrated, through simulations, that commercial building's HVAC system can provide frequency regulation by adjusting the setpoint of the duct static pressure [99]. In [37], the authors showed that up to 15% of the rated supply fans' power can be used to provide regulation reserves in the frequency range $f \in [1/(3 \text{ min}), 1/(8 \text{ sec})]$. In addition, this frequency range can be extended down to $1/(1 \text{ hr})$ if chillers are incorporated [54]. Researchers have also investigated the feasibility of using an aggregation of buildings to provide reserves: [5] studied the problem of contract design and [94] proposed a hierarchical control scheme for this scenario.

On the experimental side, there has been only a few field tests, and most of them were conducted in controlled laboratory environments with minimal uncertainties. *Maasoumy et al.* showed that variations in the setpoint of the duct static pressure can indirectly control the supply fans' power [62]. In [55], experiments were conducted in an unoccupied auditorium where filtered versions of the frequency regulation signal were tracked over 40 minute time durations using two distinct control inputs: the supply fans' speed and the setpoint of the supply air flow rate. In their work, the baseline consumption was determined *a posteriori* using a high pass filter. Nevertheless, it is desirable to determine the reserve capacity and baseline at the beginning of the regulation period in order to comply with the current AS market rules. *Vrettos et al.* proposed such a method in [93] using a MPC-based reserve scheduler. In addition, *Vrettos et al.* carried out extensive experiments in single-zone unoccupied test cells [92].

In another interesting contribution, the power consumption of electric heaters was controlled to provide frequency regulation [20]. To minimize uncertainties, ex-

periments were conducted in unoccupied rooms at nighttime. The same group of researchers then extended these results with a formulation to readjust the baseline in the intraday market to account for prediction errors, and carried out experiments during the day in occupied offices, where indoor climate is controlled using electric heaters [30]. Electric heaters have the advantage of being simpler to model compared to VAV HVAC systems and are found in many European buildings. However, since many commercial buildings in the U.S. are equipped with the latter, experimental demonstration using VAV HVAC systems is essential for their wide adoption in the U.S..

In this chapter, we demonstrate that VAV HVAC systems can be controlled to provide frequency regulation through field experiments conducted on an occupied commercial building during both daytime and nighttime. Our contributions are two-fold:

First, we aim to improve existing frequency regulation control algorithms. Unlike the frequency tracking controller in [93], which relies on an accurate supply fan model, we propose a controller that is suitable when the building is subject to larger disturbances and/or modeling uncertainties.

Second, on the experimental side, as far as we know, this is the first report where an occupied commercial building equipped with a VAV HVAC system successfully provides frequency regulation. Experiments are conducted in accordance with PJM’s certification test (40 minute duration) and tracking requirements (over multiple hours), using historical PJM regulation signals. Good tracking performance was achieved despite large disturbances such as occupancy and the use of a simple building model, which demonstrates the robustness of our method to uncertainties.

This chapter is organized as follows. We first briefly describe our control scheme in Section 3.2. Then, Section 3.3 presents the fan model, followed by Section 3.4, which describes the controller design in detail. Section 3.5 then demonstrates the performance of our controller through extensive field experiments. Finally, Section 3.6 concludes.

3.2 Problem Statement

3.2.1 PJM Regulation Market Operation [72]

In this section, we briefly describe the operation of the PJM regulation market relevant to this work, which involves the day-ahead market and the hourly scheduling process. First, resources that wish to participate in the regulation market can submit their bids, which includes capacities and offering prices, one day ahead in the day-ahead market for the entire 24 hours of the operating day. After the day-ahead market closes, PJM calculates the initial regulation schedule for each hour of the operating day, based on bids, offers, submitted schedules and predicted reserve

needs. The hourly scheduling process operates at an overlapping time frame, and allows participating resources to declare their regulation capacities and baseline consumptions, which must remain fixed for each operating hour, up to 60 minute before the beginning of the operating hour.

Assumption 1 *In this chapter, we assume that our building participates in the hourly reserve scheduling process.*

During real time operation, a frequency regulation signal is sent from PJM to each participating resource every 2 seconds. In turn, the resource is required to regulate its instantaneous electricity consumption around its baseline such that the deviation tracks the regulation signal.

3.2.2 Control Scheme

The field experiments are conducted on the fourth floor of SDH, the same building used for model identification in Chapter 2, which is a seven-floor-building equipped with a VAV HVAC system. On the control system side, SDH is equipped with the Siemens' APOGEE system – a BAS that controls the building's HVAC system and lighting based on existing control loops. The communication between the building automation devices is enabled by a building automation and control network (BACnet). We develop an external frequency regulation controller to read measurements from and send control commands to the HVAC equipment via the BACnet.

Our goal is to adjust the instantaneous power consumption of the supply fans to provide frequency regulation while maintaining the indoor temperature on the fourth floor within the comfort zone. We propose to control the supply fans' power consumption by adjusting their speed and use the airflow rates to the fourth floor for the comfort goal. The airflow rates to the rooms are chosen not to depend on the frequency regulation signal because they have very different reaction times: the regulation signal is updated every 2 seconds, however the dampers in the VAV boxes have a time constant in the order of minutes. Consequently, our controller overrides the BAS' supply fans' speed control loops and the airflow rate control loops at the VAV boxes situated on the fourth floor. All other BAS' control loops are left intact. Our proposed control architecture consists of a High Level Controller and a Low Level Controller as depicted in Figure 3.1:

- High Level Controller (HLC) - Reserve Scheduling and Room Temperature Control

A closed-loop MPC that operates every hour. Its objective is two-fold: first, determine the reserve capacity that the building can reliably offer for the next operating hour; second, calculate the baseline airflow rate to each room that ensures occupant comfort while providing this amount of reserves. Both the

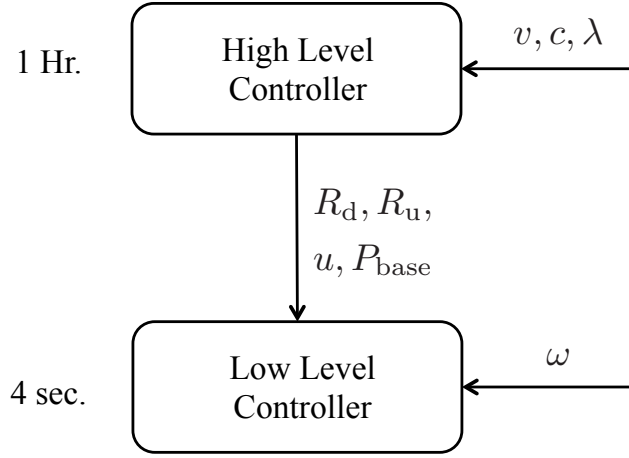


Figure 3.1: Two level control scheme for frequency regulation. See Table 3.1 for description of symbols.

capacity and the baseline are chosen to minimize electricity cost and maximize rewards from reserve provision.

- Low Level Controller (LLC) - Frequency Tracking

An improved version of the switched controller proposed in [93], which modulates the speed of the supply fan every 4 seconds so that its power consumption deviations from its baseline tracks the frequency regulation signal. It consists of a model-based feedforward controller and a Proportional Integral (PI) feedback controller. As explained in Section 3.4.2, our switched controller has a different switching condition from [93], which makes our controller suitable for scenarios where the building is subject to larger disturbances.

Note that although PJM's regulation signal is updated every 2 seconds, power consumption measurements of the supply fans in SDH are only available every 10 to 20 seconds, therefore the LLC was chosen to run at an intermediate rate of 4 seconds. This is a restriction of our particular building. Using another building whose supply fans' power measurements are updated more frequently would only improve our controller's tracking performance.

3.3 Model Identification

3.3.1 Building Model

The data-driven model identified for the fourth floor of SDH in Chapter 2 is used here in our field experiments. To recap, we model the room temperature evolution as follows:

$$x(k+1) = Ax(k) + Bu(k) + Cv(k) + q(k), \quad (3.1)$$

where $x \in \mathbb{R}^6$ represents the average temperature in each of the six zones on the fourth floor (see Figure 2.1), $u \in \mathbb{R}^6$ contains the total airflow to each zone, $v := [v_{Ta}, v_{Ts}]^\top \in \mathbb{R}^2$ is a disturbance vector that describes ambient air temperature and the HVAC system's SAT and $q \in \mathbb{R}^6$ contains internal gains due to occupancy and electric devices in each zone. Finally, $A, B \in \mathbb{R}^{6 \times 6}$ and $C \in \mathbb{R}^{6 \times 2}$ are coefficient matrices.

3.3.2 Fan Model

SDH's HVAC system contains two AHUs, each of them houses a set of supply fans that operate at the same speed at all times. In this work, we model the fans in both AHUs as a single unit, i.e., we control them to the same fan speed, we consider their total power consumption and the total air flow rate through both AHUs. With this setup, a fan model is identified from 6 weeks of one-minute resolution data collected from sMAP. The fan laws state that the airflow rate through the fan is proportional to the fan speed, and the fan power is a cubic function of its speed [67]. Figure 3.2 confirms the linear relationship between the airflow rate and the fan speed. It also shows that for the range of fan speeds used in this work (i.e., 20% to 60% of maximum fan speed), a quadratic function can describe the relationship between fan power and fan speed without significant loss of accuracy compared to a cubic function, and since a quadratic function would simplify the subsequent controller design problem, it is adopted in this work.

Let N denote the fan speed, u the total airflow through the fans and P the total fan power consumption, then the fan model is given as follows:

$$\begin{aligned} N &= f(u) = a_1 u + a_2 \\ P &= g(N) = b_1 N^2 + b_2 N + b_3 \\ P &= h(u) = c_1 u^2 + c_2 u + c_3. \end{aligned} \quad (3.2)$$

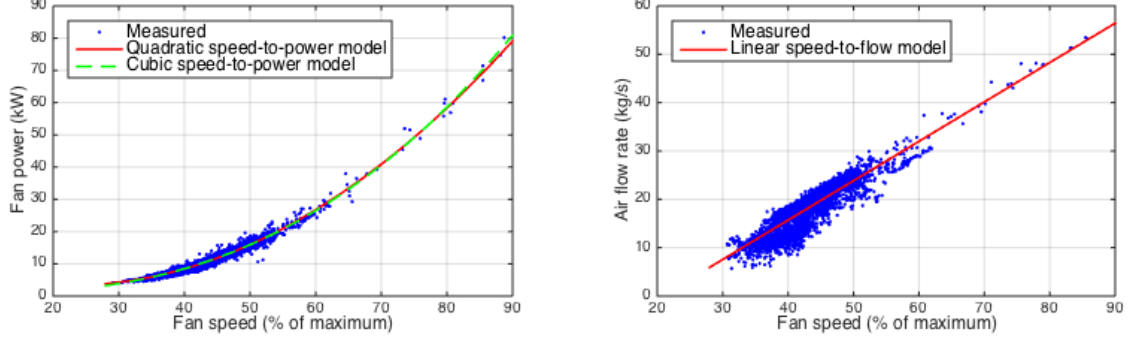


Figure 3.2: Fan measurements and identified models for fan power and air flow as functions of fan speed.

3.4 Two-Level Control Scheme

3.4.1 High Level Controller

3.4.1.1 Regulation Capacities

Let P_{base} and u_{base} represent the power consumption of the supply fans and the airflow rate through the fans at the baseline operating point. Then,

$$\begin{aligned} P_{\text{base}}(k) &= h(u_{\text{base}}(k)) \\ &= h(1^\top u(k) + v_{\text{in}}(k)), \end{aligned} \quad (3.3)$$

where $u(k) \in \mathbb{R}^6$ represents the controlled airflow rates to the six zones on the fourth floor, and $v_{\text{in}}(k) \in \mathbb{R}$ is a disturbance that represents the total airflow rate to the remaining floors. Now, define $R_{\text{u}}(k)$ and $R_{\text{d}}(k)$ as the up- and down-regulation capacities at time step k , respectively¹. In addition, let $r_{\text{u}}(k)$ and $r_{\text{d}}(k)$ denote the maximum changes in fan speed at time step k as a result of reserve provision. Then, the regulation capacities are as follows

$$\begin{aligned} R_{\text{u}}(k) &= P_{\text{base}}(k) - P_{\text{u}}(k) \\ &= h(u_{\text{base}}(k)) - g(f(u_{\text{base}}(k)) - r_{\text{u}}(k)) \\ R_{\text{d}}(k) &= P_{\text{d}}(k) - P_{\text{base}}(k) \\ &= g(f(u_{\text{base}}(k)) + r_{\text{d}}(k)) - h(u_{\text{base}}(k)), \end{aligned} \quad (3.4)$$

¹For demand resources, up-regulation requires a reduction in its power consumption and down-regulation requires an increase in consumption.

where P_u and P_d denote the fans' power consumption when providing maximum up- and down-regulation capacities, respectively.

During operating hour, the building receives a normalized real-time regulation signal $\omega \in [-1, 1]$ from PJM, which represents the requested regulation amount R as a fraction of the declared capacities R_u and R_d . In other words, the building must control its power consumption P at time k to track the following reference (desired) value:

$$\begin{aligned} P_{\text{ref}}(k) &= P_{\text{base}}(k) + R(k) \\ &= P_{\text{base}}(k) + \begin{cases} \omega(k)R_u(k) & \text{if } \omega(k) < 0 \text{ (up-regulation)} \\ \omega(k)R_d(k) & \text{if } \omega(k) \geq 0 \text{ (down-regulation).} \end{cases} \end{aligned} \quad (3.5)$$

3.4.1.2 Input Constraints

The airflow rates to the fourth floor must remain fixed for each hour in order to maintain a constant baseline, in addition to being restricted by the minimum and maximum airflow settings of the HVAC system. Note that in our experiment setup, u is unaffected by the uncertain regulation signal w .

$$u(k) = u(k + j) \text{ for all } k = 4m + 1, j = \{1, 2, 3\}, m \leq n/4 - 1, m \in \mathbb{N}, \quad (3.6)$$

$$u_{\min} \leq u(k) \leq u_{\max}. \quad (3.7)$$

Furthermore, there are minimum and maximum fan speed requirements: $N_{\min} = 20\%$ and $N_{\max} = 60\%$, where the maximum limit ensures that the HVAC system's duct static pressure remains under the maximum safe value of 2 inches water column, and the minimum limit ensures that the supply fans have sufficient power to drive the supply air throughout the building:

$$N_{\min} \leq N(k) \leq N_{\max} \text{ for all } \omega(k) \in [-1, 1], \quad (3.8)$$

where $N(k)$ is the fan speed at time step k . Since ω is unknown at scheduling time, the above constraint must hold for any ω . The fan model states that N is given by the inverse of the speed-to-power function, $N = g^{-1}(P)$, which is usually complex, however, it has a simple form at the boundary values of ω :

$$N(k) = \begin{cases} N_{\text{base}}(k) - r_u(k) & \text{if } \omega(k) = -1 \\ N_{\text{base}}(k) + r_d(k) & \text{if } \omega(k) = 1. \end{cases} \quad (3.9)$$

Thus, the above constraints can be satisfied by imposing the following robustified version of (3.8)

$$N_{\min} + r_u \leq N_{\text{base}}(k) \leq N_{\max} - r_d. \quad (3.10)$$

Although [93] showed that the energy content of ω over a 15 minute interval is typically limited and $\omega \in [-\omega_{\text{lim}}, \omega_{\text{lim}}]$ most of the time, where $0 < \omega_{\text{lim}} < 1$; we choose to deal with the uncertainty introduced by ω in a robust fashion, by accounting for the worst case, i.e., $\omega = \pm 1$. This way, the building need not make any assumptions on the statistical properties of ω , and it has the added advantage of building additional robustness to modeling and forecast uncertainties for an occupied building.

3.4.1.3 Output Constraints

The indoor temperature x should be kept within the time varying comfort zone $[x_{\min}(k), x_{\max}(k)]$ for all k :

$$x_{\min}(k) \leq x(k) \leq x_{\max}(k). \quad (3.11)$$

Note that x is unaffected by the unknown signal ω since u is independent of ω .

The regulation capacities are required to be fixed for each hour which translates to the following constraint:

$$\begin{aligned} R_u(k) = R_u(k+j) \text{ and } R_d(k) = R_d(k+j) \\ \text{for all } k = 4m+1, j = \{1, 2, 3\}, m \leq n/4-1, m \in \mathbb{N}. \end{aligned} \quad (3.12)$$

In addition, these capacities must be non-negative. Because fan power is a non-decreasing function of its speed in the operating range $[N_{\min}, N_{\max}]$, the following condition ensures $R_u(k) \geq 0$ and $R_d(k) \geq 0$:

$$r_u(k) \geq 0, r_d(k) \geq 0. \quad (3.13)$$

3.4.1.4 Cost Function

The objective of the HLC is to choose the HVAC's operating point to minimize the cost of electricity consumption and maximize rewards from reserve provision.

Assumption 2 *Electricity cost is calculated based on the baseline consumption.*

Assumption 3 *The payment for both up- and down-regulations are the same.*

Then, the building's objective is to minimize the following cost function

$$c(k)P_{\text{base}}(k) - \lambda(k)(R_u(k) + R_d(k)), \quad (3.14)$$

where c denotes the unit electricity cost and λ denotes the reward for providing regulation service.

3.4.1.5 Robust Optimization Problem

We introduce the following robust optimization problem:

$$\begin{aligned}
& \underset{u(k), r_u(k), r_d(k)}{\text{minimize}} && \sum_{k=1}^n c(k)P_{\text{base}}(k) - \lambda(k)(R_u(k) + R_d(k)) \\
& \text{subject to} && u_{\min} \leq u(k) \leq u_{\max} \\
& && N_{\min} + r_u \leq N_{\text{base}}(k) \leq N_{\max} - r_d \\
& && x(k) = Ax(k) + Bu(k) + Cv(k) + q(k) \\
& && x_{\min}(k) \leq x(k) \leq x_{\max}(k) \\
& && r_u(k) \geq 0, \quad r_d(k) \geq 0 \\
& && \text{conditions (3.6) and (3.12) for constant} \\
& && \text{hourly baseline and regulation capacities.}
\end{aligned} \tag{3.15}$$

(3.15) is a deterministic nonlinear optimization, with non-convex quadratic cost function and linear equality and inequality constraints. Despite its non-convexity, it can be solved in Matlab using YALMIP [47] and the *fmincon* solver in less than one second, due to its small size. The outcome is the baseline operating point for the HVAC system, and the up- and down-regulation capacities $R_u(k)$ and $R_d(k)$ for each time step in the scheduling horizon $k = 1, \dots, n$.

3.4.1.6 Disturbance Approximation and Forecast

The building's dynamics are subject to the following disturbances: ambient temperature v_{Ta} , SAT v_{Ts} and internal gains due to occupancy q . Ambient temperature forecasts are obtained from the publicly available database of *darksky.net* [59]. Analysis of historical data indicates that SAT rarely varies, as a result, it is measured at each time step k and assumed constant for the HLC horizon, i.e., 1 hour for the certification experiment and 4 hours for the tracking experiment. The internal gains q is first estimated at each time step k from the building model and real-time measurements, and then assumed constant for the HLC horizon. In addition, the cost function (3.14) is affected by v_{in} , the total airflow rate to other floors of SDH, through (3.3) and (3.4). This is also measured at each time step k and assumed constant for the HLC horizon.

3.4.2 Low Level Controller

The Low Level Controller (LLC)'s responsibility is to vary the fan power P to track the reference P_{ref} . Our approach is different from [63] which used the frequency of the variable frequency drive as the control input, and from [55] where control inputs, acting as disturbances, are superimposed onto HVAC's existing control loop commands.

We improve on the switched controller proposed in [93], so that it is suitable for buildings subjected to large disturbances where an accurate fan model is not available. It consists of two sub-controllers. Controller 1 is a model-based feedforward controller which uses the fan model from (3.2) to determine the fan speed required for a given $P_{\text{ref}}(k)$, i.e., $N(k) = g^{-1}(P_{\text{ref}}(k))$. This feedforward controller has a fast response and thus, is best suited when there is a large change in the reference regulation signal, i.e., $|P_{\text{ref}}(k) - P_{\text{ref}}(k-1)| > \epsilon$ where ϵ is a user defined threshold value. On the other hand, Controller 1 has non zero steady state error due to inaccuracies in the fan model. Therefore, Controller 2 – a PI controller, is used to reduce any steady state error from Controller 1. The implementation of the LLC is presented in Algorithm 1. Step 13 is the discrete implementation of the PI controller, where Δt is the discretization time step (4 seconds in our case). In step 15, the fan speed is capped between N_{min} and N_{max} to satisfy constraint (3.8).

Algorithm 1 Low Level Controller

```

1: Initialize previous tracking error  $e(k-1) = 0$ , reference regulation signal  $P_{\text{ref}}(k-1) = 0$  and fan speed  $N(k-1)$ 
2: while experiment is running do
3:   Compute baseline fan power  $P_{\text{base}}(k) = h(u_{\text{base}}(k))$ 
4:   Compute reference (desired) fan power  $P_{\text{ref}}(k) = P_{\text{base}}(k) + \omega(k)R_u(k)$  if  $\omega(k) < 0$  and  $P_{\text{ref}}(k) = P_{\text{base}}(k) + \omega(k)R_d(k)$  if  $\omega(k) \geq 0$ 
5:   Measure actual fan power  $P(k)$ 
6:   Compute current tracking error  $e(k) = P_{\text{ref}}(k) - P(k)$ 
7:   if  $|P_{\text{ref}}(k) - P_{\text{ref}}(k-1)| > \epsilon$  then
8:     Use model-based controller:
9:      $N(k) = g^{-1}(P_{\text{ref}}(k))$ 
10:  else
11:    Use PI controller:
12:     $N(k) = N(k-1) + K_P(e(k) - e(k-1)) + \frac{K_P}{K_I}e(k) \cdot \Delta t$ 
13:  end if
14:  Cap fan speed:  $N(k) = \max(\min(N(k), N_{\text{max}}), N_{\text{min}})$ 
15:  Update variables:  $e(k-1) \leftarrow e(k)$ ,  $P_{\text{ref}}(k-1) \leftarrow P_{\text{ref}}(k)$  and  $N(k-1) \leftarrow N(k)$ 
16: end while

```

The improved performance of our switched controller compared to that in [93] is a result of the new switching condition. In [93], the LLC switches between its sub-controllers based on whether the absolute tracking error is greater than a threshold $\bar{\epsilon}$, i.e., $|e(k)| > \bar{\epsilon}$. To see why this controller may fail when an accurate fan model is not available, consider a step change in the reference signal P_{ref} at time step k such that $|e(k)| > \bar{\epsilon}$. According to the switching condition in [93], Controller 1 would become active. If the fan model is accurate enough such that the fan speed given by the inverse fan model is able to reduce the tracking error at the next time step $k+1$

to $|e(k+1)| \leq \bar{\epsilon}$, then Controller 2 would take over and continue to decrease the tracking error from time $k+1$ onwards. On the other hand, if the fan model is not accurate enough such that $|e(k+1)| > \bar{\epsilon}$, then the switched controller in [93] is stuck in Controller 1 with a constant fan speed $N(k) = g^{-1}(P_{\text{ref}}(k))$, and consequently, a constant tracking error whose absolute value is greater than $\bar{\epsilon}$. In practice, accurate fan models are often not available for commercial buildings like SDH which are subject to large disturbances and uncertainties. Therefore, larger values of $\bar{\epsilon}$ are needed to avoid the above problem. We found that for SDH, the required value of $\bar{\epsilon}$ would be so large that the switched controller essentially acts as a simple PI controller, and a PI controller alone is unable to achieve the fast response necessary for a good tracking performance.

Our proposed switched controller overcomes the above problem and is therefore suitable for buildings where an accurate fan model is not available. Consider the above scenario again, $|P_{\text{ref}}(k+1) - P_{\text{ref}}(k)| = 0 \leq \epsilon$ independent of the fan model, therefore Controller 2 is activated at time step $k+1$ and continues to reduce the tracking error.

3.4.2.1 PI Controller Tuning

The proportional K_P and integral gains K_I of the PI controller are calculated using the Open Loop Ziegler Nichols method [71]. Open loop responses of the fan power to step changes in its speed were recorded. The delay time and time constant of the response were used to compute the gains K_P and K_I . These values served as initial guesses, which we fine tuned later through trial and error. The final gain values are: $K_P = 0.3$, $K_I = 6.8$.

Note that the Closed Loop Ziegler Nichols tuning method used in [93] may be unsafe for a regular building, because it requires increasing the proportional gain K_P until the fan power exhibits sustained oscillations, i.e., the system is marginally stable. Also note that a single PI controller for the entire operating range of the supply fans was found to be sufficient to achieve good tracking performance.

3.5 Experimental Results

3.5.1 Communication Architecture

The communication architecture is shown in Figure 3.3. The HLC is implemented in Matlab on a standard laptop and the LLC is implemented in Python on a server located in SDH. The Matlab and Python scripts communicate asynchronously via local port forwarding and TCP/IP. All requests are initiated from the Matlab script and forwarded to the connected port on the laptop, while the Python script continuously listens to the corresponding port on the server and responds to any received requests.

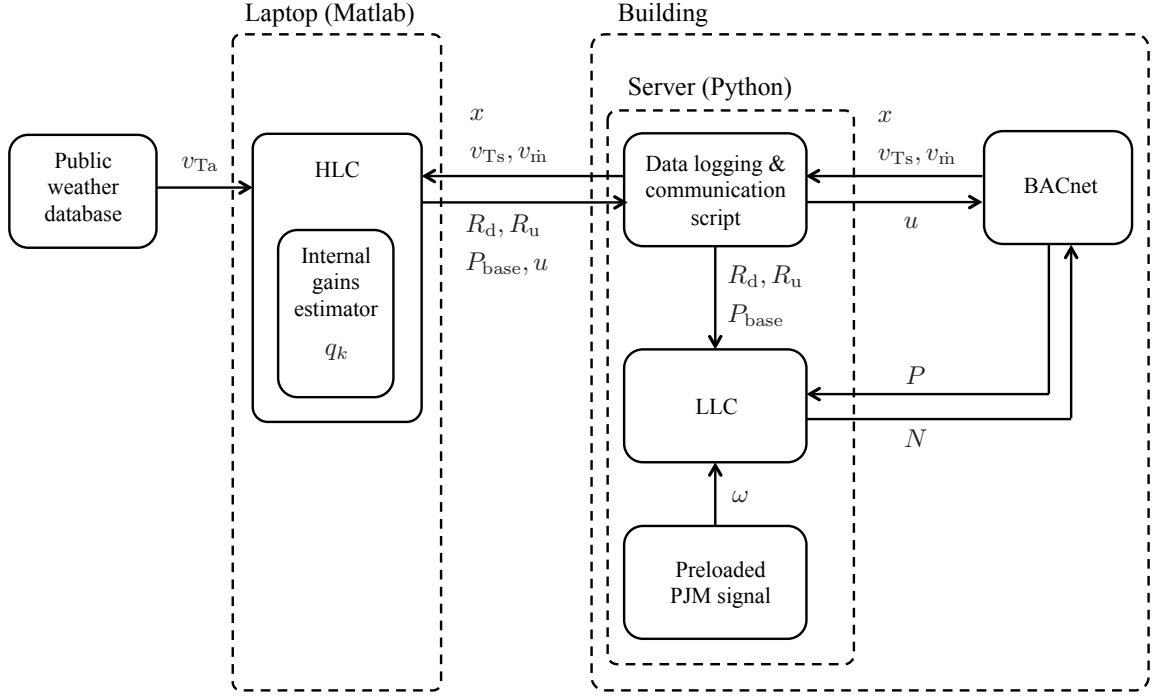


Figure 3.3: Communication architecture. See Table 3.1 for description of symbols.

The HLC queries the publicly available database of *darksky.net* to obtain weather forecasts, and collects building measurements by sending read requests to the Python script, which then directly queries the BACnet to obtain these measurements. At the same time, the HLC computes the regulation capacities, baseline and optimal airflow rate setpoints and sends them as write requests to the Python script. The script then updates the values of its local copy of the regulation capacities and baseline, and adjusts the appropriate actuator setpoints via BACnet. Finally, the LLC reads the preloaded PJM regulation signal, and determines and adjusts the fan speed setpoint using read and write requests via the BACnet.

3.5.2 Performance Metric

PJM's composite performance score S_{comp} is used to evaluate our controller's performance in tracking the regulation signal. This score consists of three equally weighted components: a correlation score S_c , a delay score S_d and a precision score S_p , defined as in [78]:

$$\begin{aligned}
S_c &= \max_{t \in [0, 5\text{min}]} (\sigma_t) \\
S_d &= \frac{1}{5\text{min}} \left(5\text{min} - \frac{\max(t^* - 10\text{sec}, 0)}{60\text{sec}} \right) \\
&\quad \text{where } t^* = \arg \max_{t \in [0, 5\text{min}]} (\sigma_t) \\
S_p &= 1 - \frac{1}{n_h} \sum_{k=1}^{n_h} \frac{|P_{\text{ref}}(k) - P(k)|}{0.5 \cdot (R_{d,h} + R_{u,h})} \\
S_{\text{comp}} &= \frac{1}{3} \cdot (S_c + S_d + S_p).
\end{aligned} \tag{3.16}$$

In (3.16), σ_t represents the correlation between the reference signal $P_{\text{ref}}(k)$ and the fans' actual power consumption delayed by t seconds $P(k+t)$. In other words, S_c measures the maximum correlation between the regulation signal and the building's response signal within each 5 minute window. S_d represents the response delay when maximum correlation occurs, with a "free" 10 second allowance, and finally S_p measures the average tracking error scaled by the building's regulation capacity offered for that hour h . All scores take values between 0 and 1, with 1 being a perfect score. Following PJM's rule, we generate S_c and S_d every 10 seconds and calculate S_p once per hour. The performance score S_{comp} is then computed every hour by averaging S_c and S_d scores for the hour and using S_p for that hour.

3.5.3 Certification Experiment

Any resource that intends to participate in PJM's regulation market must first pass the certification test, which is run during a continuous 40-minute period, using the test signal published on PJM's website [82]. The test is scored using S_{comp} evaluated on the entire 40-minute test period, and a resource is certified only after it achieves three consecutive scores of 0.75 or above. In addition, both the baseline consumption and regulation capacity must be declared before the test begins and remain constant throughout the test.

We carried out four certification tests at various hours of the day from November 27 (Sunday) to 28 (Monday) 2016 and achieved S_{comp} values around 0.9 in all tests (Table 3.2). This demonstrates that our controller's performance is robust to disturbances such as weather and occupancy. Figure 3.4 shows the desired reference signal P_{ref} , the fan's actual power consumption P and the percentage tracking error defined as $(P_{\text{ref}} - P)/P_{\text{ref}} \cdot 100$ for the test conducted on November 28 starting at 10 a.m.. An absolute tracking error of less than 28% is observed during 90% of the test period.

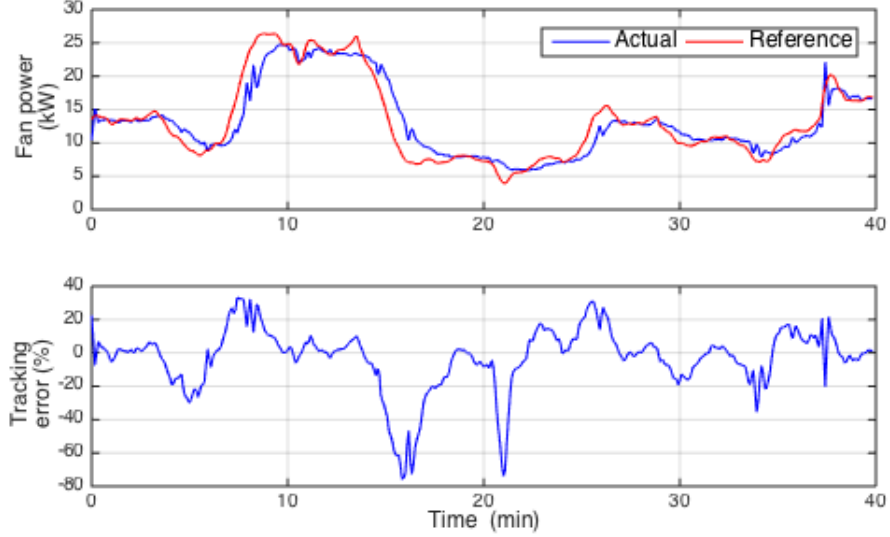


Figure 3.4: Certification experiment results: power and tracking error.

3.5.4 Tracking Experiment

In this section, we present the results from the tracking experiment conducted from 13:00 to 17:00 on Nov 29, 2016, which uses the historic PJM RegD signal recorded from 13:00 to 17:00 on July 1, 2016 [82] as the reference frequency regulation signal.

After successful certification, each resource must maintain a historic performance score S_{comp} of above 40% to continue to provide regulation services. In addition, a resource must achieve an average hourly score of at least 50% to receive payments for offering regulation capacity. Table 3.3 shows the S_{comp} scores calculated for each hour, as well as the average scores, of our tracking experiment. Our controller scored S_{comp} values well above both thresholds during the field experiment, demonstrating good tracking performance.

Figure 3.5 shows that the actual fan power closely tracks the reference power signal and indeed, an absolute tracking error of less than 16% is achieved 90% of the time. The error is greatest at the start of the experiment as the fans switch from normal operation to frequency regulation mode, and decays as the experiment progresses. The fans offer a constant total regulation capacity of 23.3 kW every hour, but the split between up- and down-regulation capacities varies hourly as the building’s baseline consumption changes. Finally, we confirm that the duct static pressure remains below the maximum limit of 2 inches water column throughout the experiment.

We present the zone temperatures and airflow rates from the experiment in Figure 3.6, and note the following observations. First, all zone temperatures are

maintained within the comfort bounds and flow rates are kept within the permissible ranges. Second, to minimize electricity cost, flow rates are kept at their minima unless continued supply of minimum airflow to a zone is predicted to cause the zone temperature to exceed its maximum limit within the prediction horizon. For example, flow rates are increased in the second and third hours in the West zone to maintain occupant comfort. Third, observe that temperatures decrease during the third hour in the West zone and the second hour in the South zone, which indicates that the rooms are overcooled, i.e., supply air's flow rates are more than the minimum required to maintain zone temperatures within the comfort range. This is likely due to disturbance prediction errors.

3.6 Conclusion

In this chapter, we demonstrate experimentally that commercial buildings equipped with VAV HVAC systems can provide frequency regulation, by varying the power consumption of the supply fans. To do so, we improve on existing frequency regulation control algorithms and propose a two-level control scheme that is suitable for buildings subjected to larger disturbances and/or modeling uncertainties. Then, we demonstrate our controller's performance through numerous field experiments conducted in an occupied commercial building, using archived PJM frequency regulation signals.

Symbol	Description
R_u, R_d	Up- and down-regulation capacities
P_{base}, P	Baseline and actual fan power consumption
N_{base}, N	Baseline and actual fan speed
x	Zone temperature
u	Control input
v	Disturbance
q	Internal gains
w	Normalized frequency regulation signal
c	Unit electricity cost
λ	Reward for reserve provision

Table 3.1: Table of notation.

Test date	Test start time	S_c	S_d	S_p	S_{comp}
Nov 27	22 hours	0.90	0.90	0.88	0.89
Nov 28	10 hours	0.93	0.93	0.88	0.91
Nov 28	13 hours	0.92	0.90	0.89	0.90
Nov 28	16 hours	0.93	0.92	0.89	0.91

Table 3.2: PJM performance scores for certification tests.

	S_c	S_d	S_p	S_{comp}
1 st hour	0.94	0.91	0.99	0.95
2 nd hour	0.72	0.52	0.99	0.74
3 rd hour	0.63	0.63	0.99	0.75
4 th hour	0.62	0.31	0.99	0.64
Mean	0.73	0.59	0.99	0.77

Table 3.3: PJM performance scores for tracking experiment.

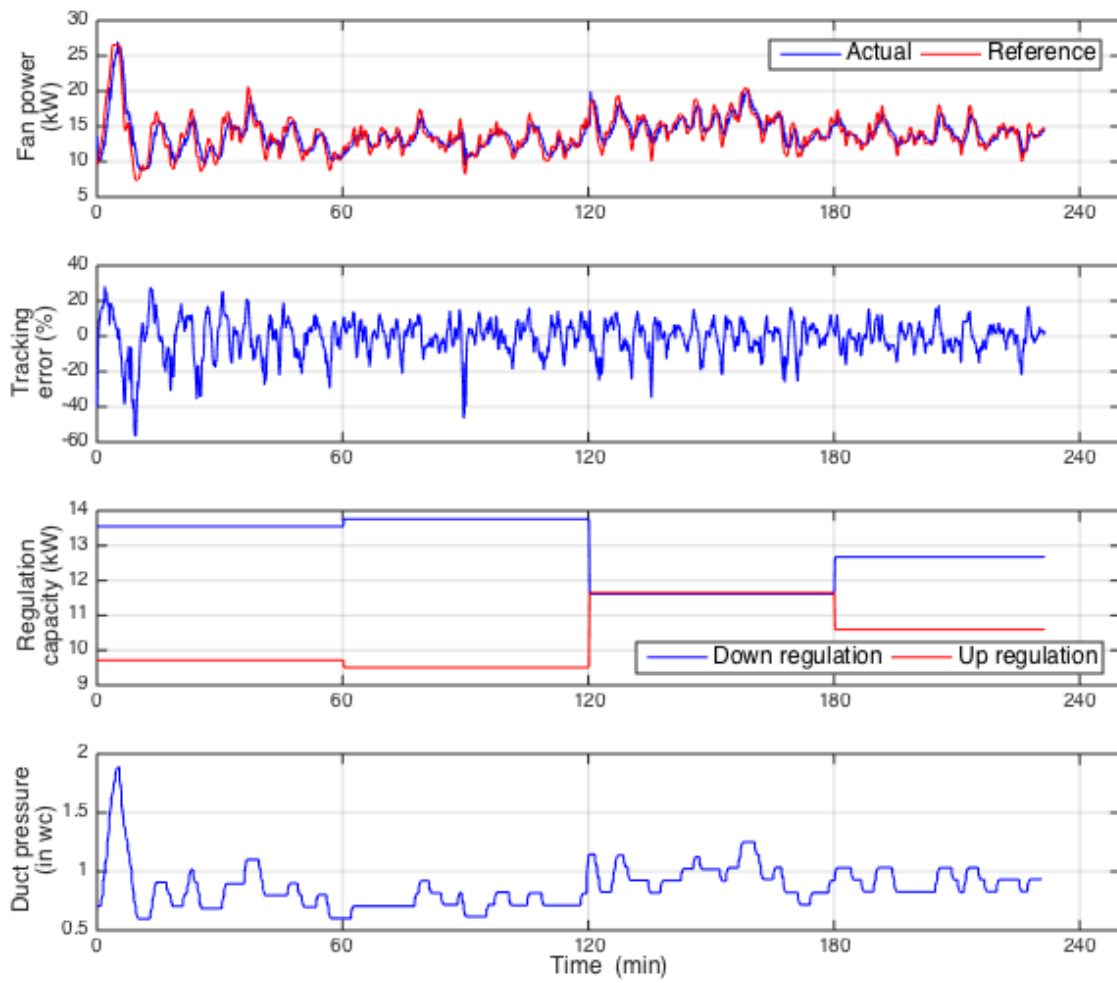


Figure 3.5: Tracking experiment results: power, tracking error, regulation capacities and duct pressure.

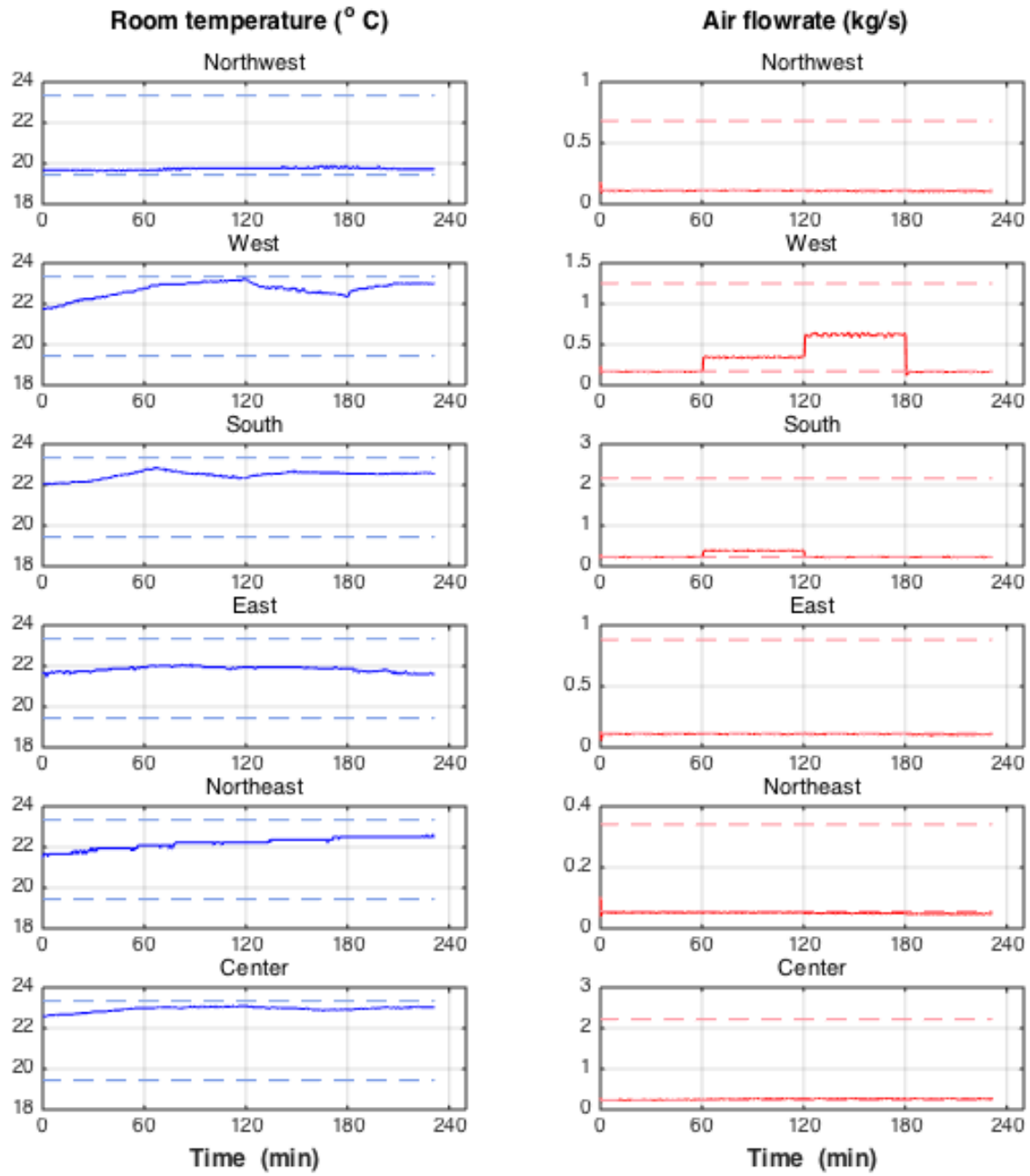


Figure 3.6: Tracking experiment results: room temperatures and airflow rates. Solid lines are actual values, dashed lines are maximum and minimum limits.

Part 2

Secure Estimation under Cyber Attacks

Chapter 4

Secure Estimation for Linear Systems

In this Chapter, we develop a secure estimator for a general linear system and demonstrate its effectiveness through an example of an unmanned aerial vehicle (UAV) under cyber attack. In the next Chapter, we extend these results to two classes of nonlinear systems and apply the secure estimator to the nonlinear power system.



CPSs are found in many applications such as power networks, manufacturing processes, air and ground transportation systems. They consist of physical components such as actuators, sensors and controllers that communicate with each other over a network [48]. For example, UAVs may obtain position measurements from a GPS or communicate with a remote control center (RCC). Although communication networks are often protected by security measures, cyber attacks can still take place when a malicious attacker obtains unauthorized access, launching jamming attacks [28], or spoofing sensor readings and sending erroneous control signals to actuators [68]. For CPS, cyber attacks not only compromise information but can also cause damage in the physical process, ranging from power systems [57, 89] to UAVs [43]. This presents new challenges and thus demands new strategies and algorithms [11].

Maintaining security of these systems under cyber attacks is an important and challenging task, since these attacks can be erratic and thus difficult to model. Secure estimation problems study how to estimate the true system states when measurements are corrupted and/or control inputs are compromised by attackers. In designing such estimators, it is desirable to make as few assumptions about the attackers as possible. This is because it is very difficult, if not impossible, to predict the behavior of attackers, and when an attack signal violates the assumptions of a secure estimator, then this estimator would fail to detect the attack. [24] proposed a novel secure estimation method assuming that attack signals can be arbitrary and unbounded. However, one limitation of their proposed estimator is that the set of attacked sensors (sensors, controllers) is assumed to be fixed. In this chapter, we extend these results to scenarios

in which the set of attacked sensors can change over time. We formulate this secure estimation problem into the classical error correction problem [10] and we show that accurate estimation can be guaranteed. Furthermore, we propose a combined secure estimation method with our proposed secure estimator and the Kalman Filter (KF) for improved practical performance. Finally, we demonstrate the performance of our method through simulations of two scenarios where a UAV is under cyber attack.

This chapter is an adaptation of the paper in [43].

4.1 Introduction

Researchers have studied various approaches to securing CPS. Each of them relies on specific assumptions about attackers' strategies and it is rarely the case, if not impossible, that one estimator/detector can protect against all possible attacks. For example, [50, 53, 57, 89] studied optimal attack strategies for different control systems and applications. From the controller's point of view, [64, 76] assumed that the attack signal would follow certain probabilistic distributions and then designed filters to detect these attacks. In [25, 33, 34, 65, 80], the authors used the game theory framework, where the controller and attacker are players with competing goals in a game. Attackers are assumed to adopt specific strategies that maximize a certain cost and the controller or estimator is designed to minimize such a cost. Finally, the authors of [52] proposed a hybrid controller, where each constituent controller protects the system against a specific type of attack.

More recently, Fawzi *et al.* studied secure estimation of a discrete time linear time invariant (LTI) system and proposed in [24] a secure estimation method for arbitrary attacks. Later, [73] and [83] extended this work by relaxing the assumption of having an exact system model and proposed an Satisfiability Modulo Theory (SMT)-based observer that handles large systems with thousands of sensors. One limiting assumption of [24], [73] and [83] is that the set of attacked sensors is fixed and can not change over time. Therefore, if a malicious attacker is aware of this assumption, then she can exploit this weakness and attack different sensors at different time steps so that such an estimator would fail to detect the presence of the attack.

In this chapter, we focus on sensor attacks on CPS and attempt to design a secure estimator for LTI systems based on as few assumptions about the attacker as possible. First, we do not assume that the attack signals follow any stochastic distributions, and thus our proposed estimator works for arbitrary and unbounded attacks. Second, we allow the set of attacked sensors to change over time. The only assumption we make is that the number of attacked sensors is sparse. Such attacks are found in many practical situations, and can be launched in both the cyber and the physical domain. For example, security studies on the current traffic infrastructure [27] demonstrated that once a cyber attacker gains access to the traffic network at a single point, the

attacker can send commands to any traffic intersection in the network. In other words, the attacker can freely attack a different set of traffic signals (sensors) at any time. Indeed, an attacker who desires to travel through a set of roads as fast as possible would attack different traffic lights to always give herself green lights as she moves through the road network. [56] describes a physical attack on a power system where a changing set of attacked sensors is desirable: in particular, the authors designed a multi-switch attack, in which different switches in a power network are attacked at different times, in order to lead to stealthy and wide-scale cascading failures in the power system.

We formulate this secure estimation problem into the classical error correction problem, from which we propose an l_1 -optimization based estimator that is computationally efficient. In addition, we prove the maximum number of sensor attacks that can be corrected with our estimator and propose to use pole placement techniques to design a feedback controller such that the resulting secure estimator can guarantee accurate estimation. Finally, to improve the estimator's practical performance, we propose to combine our secure estimator with a KF, where the KF serves to filter out both occasional estimation attacks by the secure estimator and noisy measurements, and we demonstrate the effectiveness of the combined estimator using two examples of UAVs under adversarial cyber attacks.

This chapter is organized as follows. A review of compressed sensing and error correction is given in Section 4.2, followed by Section 4.3 which proposes a secure state estimator for an LTI system under sensor attack where the attacked sensors can change over time. Then, Section 4.4 describes how pole-placement can be used to design a good secure estimator and how it can be combined with a KF to improve its practical performance. Finally we demonstrate the performance of the combined secure estimator through two numerical examples of UAVs subject to adversarial attacks in Section 4.5.

4.2 Review of Classical Error Correction

4.2.1 Compressed Sensing

Sparse solutions $x \in \mathbb{R}^n$, are sought to the following problem:

$$\min_x \|x\|_0 \text{ subject to } b = Ax \quad (4.1)$$

where $b \in \mathbb{R}^m$ are the measurements, and $A \in \mathbb{R}^{m \times n}$ ($m \ll n$) is a sensing matrix. $\|x\|_0$ denotes the number of nonzero elements of x . The following lemma provides a sufficient condition for a unique solution to (4.1).

Lemma 1 ([40]) *If the sparsest solution to (4.1) has $\|x\|_0 = q$ and $m \geq 2q$ and all subsets of $2q$ columns of A are full rank, then the solution is unique.*

Proof: Suppose the solution is not unique. Therefore, there exists $x_1 \neq x_2$ such that $Ax_1 = b$ and $Ax_2 = b$ where $\|x_1\|_0 = \|x_2\|_0 = q$. Then, $A(x_1 - x_2) = 0$ and $x_1 - x_2 \neq 0$. Since $\|x_1 - x_2\|_0 \leq 2q$ and all $2q$ columns of A are full rank (i.e., linearly independent), it is impossible to have $x_1 - x_2 \neq 0$ that satisfies $A(x_1 - x_2) = 0$. This contradicts the assumption. ■

Remark 1 (*Measurement Noise*) In practice, the measurements are noisy so one cannot assume that the Ax term in (4.1) is known with arbitrary precision. More appropriately, we need to assume that one is given noisy measurements, i.e., $b = Ax + \epsilon$, where ϵ represents measurement noise. In [9], the authors prove that one can recover approximately sparse signals with an error at most proportional to the noise level. An alternative is to combine secure estimation with a KF to improve the secure estimator's performance for noisy measurements as we propose later in this chapter: the KF filters out both occasional estimation errors by the secure estimator and noisy measurements.

4.2.2 The Error Correction Problem [10]

Consider the classical error correction problem: $y = Cx + e$ where $C \in \mathbb{R}^{l \times n}$ is a coding matrix ($l > n$) and assumed to be full rank. We wish to recover the input vector $x \in \mathbb{R}^n$ from corrupted measurements y . Here, e is an arbitrary and unknown sparse error vector. To reconstruct x , note that it is obviously sufficient to reconstruct the vector e since knowledge of $Cx + e$ together with e gives Cx , and consequently x since C has full rank [10]. In [10], the authors construct a matrix F which annihilates C on the left, i.e., $FCx = 0$ for all x . Then, they apply F to the output y and obtain

$$\tilde{y} = F(Cx + e) = Fe. \quad (4.2)$$

Thus, the decoding problem can be reduced to that of reconstructing a sparse vector e from the observations $\tilde{y} = Fe$. Therefore, by Lemma 1, if all subsets of $2q$ columns of F are full rank, then we can reconstruct any e such that $\|e\|_0 \leq q$.

4.3 Secure Estimation

4.3.1 Problem Formulation

Consider the LTI system as follows:

$$\begin{aligned} x(k+1) &= A_o x(k) + Bu(k) \\ y(k) &= Cx(k) + e(k), \end{aligned} \quad (4.3)$$

where $x(k) \in \mathbb{R}^n$, $y(k) \in \mathbb{R}^p$ and $u(k) \in \mathbb{R}^m$ are the states, measurements and control inputs at time step k . $e(k) \in \mathbb{R}^p$ represents the attack signal at time k . Our goal is to

reconstruct the initial state $x(0)$ of the plant from the corrupted observations $y(k)$'s where $k = 0, \dots, T - 1$.

The attack vector $e(k)$ is such that if the i -th sensor is attacked at time k , then $e_i(k)$, the i -th element of $e(k)$ is nonzero, otherwise $e_i(k) = 0$. We assume that the attack signal can be arbitrary and unbounded. In addition, we assume that the set of attacked sensors can change over time. As illustrated by the following example, if two sensors are attacked at each time step, we can have sensors 1 and 3 attacked at time step 0, sensors 2 and 3 attacked at time 1, and so on:

$$[e(0) \mid e(1) \mid \dots] = \begin{bmatrix} * & 0 & * & \dots \\ 0 & * & 0 & \dots \\ * & * & 0 & \dots \\ 0 & 0 & * & \dots \end{bmatrix},$$

where $*$ denotes a nonzero component (i.e., an attack or corruption).

Furthermore, assume that a local control loop implements secure state feedback and is not subject to attack: $u(k) = Gx(k)$. This represents the following practical scenario: a physical system possesses a local control loop that has direct access to the state of the plant and can control the evolution of the physical system. This is reasonable if the sensors are connected to the local controller through a wired link that is not subject to external attacks. Also, as part of the overall plant, a higher-level supervisory and monitoring system receives measurements from the sensors through wireless and vulnerable communication links that are subject to attacks [24]. A concrete example is a UAV that uses measurements from onboard, hardwired sensors such as an Inertial Measurement Unit (IMU) for autopilot and trajectory following (i.e., secure local control loop), and communicates wirelessly with a remote control center (i.e., vulnerable link subject to attacks). The resulting closed loop system is:

$$\begin{aligned} x(k+1) &= Ax(k) \\ y(k) &= Cx(k) + e(k), \end{aligned} \tag{4.4}$$

where the closed loop system matrix $A = A_o + BG$.

Finally, we define the number of correctable attacks as follows:

Definition 4.3.1 *When the set of attacked sensors/nodes can change over time, q errors are correctable after T steps by the estimator/decoder $\mathcal{D} : (\mathbb{R}^p)^T \rightarrow \mathbb{R}^n$ if for any $x(0) \in \mathbb{R}^n$ and any sequence of vectors $e(0), \dots, e(T-1)$ in \mathbb{R}^p such that $|\text{supp}(e(k))| \leq q^1$, we have $\mathcal{D}(y(0), \dots, y(T-1)) = x(0)$ where $y(k) = CA^k x(0) + e(k)$ for $k = 0, \dots, T-1$.*

¹ $|\text{supp}(x)|$ denotes the support of vector x , i.e., the number of nonzero components in x . If f is any real-valued or vector-valued function on a topological space X , the support of f , denoted by $\text{supp}(f)$, is the closure of the set points where f is nonzero: $\text{supp}(f) = \{x \in X \mid f(x) \neq 0\}$.

4.3.2 Methodology

Let $E_{q,T}$ denote the set of error vectors $[e(0); \dots; e(T-1)] \in \mathbb{R}^{p \cdot T}$ where each $e(k)$ satisfies $\|e(k)\|_0 \leq q \leq p$.

$$\begin{aligned} Y &\triangleq \begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(T-1) \end{bmatrix} = \begin{bmatrix} Cx(0) + e(0) \\ CAx(0) + e(1) \\ \vdots \\ CA^{T-1}x(0) + e(T-1) \end{bmatrix} \\ &= \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{T-1} \end{bmatrix} x(0) + E_{q,T} \triangleq \Phi x(0) + E_{q,T} \end{aligned} \quad (4.5)$$

where $Y \in \mathbb{R}^{p \cdot T}$ is a collection of corrupted measurements over T time steps and $\Phi \in \mathbb{R}^{p \cdot T \times n}$ represents an observability-like matrix of the system. Here, we need to assume that $\text{rank}(\Phi) = n$; otherwise, the system is unobservable and we cannot determine $x(0)$ even if there is no attack (i.e., $E_{q,T} = 0$).

Inspired by the error correction techniques proposed in [10] and [40], we first determine the error vector $E_{q,T}$, and then solve for $x(0)$. Consider the QR decomposition of $\Phi \in \mathbb{R}^{p \cdot T \times n}$,

$$\Phi = [Q_1 \quad Q_2] \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = Q_1 R_1 \quad (4.6)$$

where $[Q_1 \quad Q_2] \in \mathbb{R}^{p \cdot T \times p \cdot T}$ is orthogonal, $Q_1 \in \mathbb{R}^{p \cdot T \times n}$, $Q_2 \in \mathbb{R}^{p \cdot T \times (p \cdot T - n)}$, and $R_1 \in \mathbb{R}^{n \times n}$ is a rank- n upper triangular matrix. Pre-multiplying (4.5) by $[Q_1 \quad Q_2]^\top$ gives:

$$\begin{bmatrix} Q_1^\top \\ Q_2^\top \end{bmatrix} Y = \begin{bmatrix} R_1 \\ 0 \end{bmatrix} x(0) + \begin{bmatrix} Q_1^\top \\ Q_2^\top \end{bmatrix} E_{q,T}. \quad (4.7)$$

We can compute $E_{q,T}$ by using the second block row:

$$\tilde{Y} \triangleq Q_2^\top Y = Q_2^\top E_{q,T} \quad (4.8)$$

where $Q_2^\top \in \mathbb{R}^{(p \cdot T - n) \times p \cdot T}$. From Lemma 1, (4.8) has a unique, s -sparse solution (where $s \leq q \cdot T$) if all subsets of $2s$ columns (at most $2q \cdot T$ columns) of Q_2^\top are full rank. Clearly, this is a reasonable assumption if $(p \cdot T - n) \geq 2q \cdot T$. Therefore, we consider solving the following l_1 -minimization problem:

$$\hat{E}_{q,T} = \arg \min_E \|E\|_{l_1} \quad \text{s.t.} \quad \tilde{Y} = Q_2^\top E, \quad (4.9)$$

where $\|E\|_{l_1} := \sum_{i=1}^{p \cdot T} |E_i|$. Now, given the vector $\hat{E}_{q,T}$, we can compute $x(0)$ from the first block row of (4.7) as follows:

$$x(0) = R_1^{-1} Q_1^\top (Y - \hat{E}_{q,T}) \quad (4.10)$$

The following lemma provides the conditions under which the solution to (4.10) exists and is unique.

Lemma 2 *$x(0)$ is the unique solution if $|\text{supp}(\Phi z)| > 2s = 2(q \cdot T)$ for all $z \in \mathbb{R}^n \setminus \{0\}$.*

Proof: We first prove the claim C1: if $|\text{supp}(\Phi z)| > 2s = 2(q \cdot T)$ for all $z \in \mathbb{R}^n \setminus \{0\}$ then all subsets of $2s$ columns of Q_2^\top are full rank. Then by Lemma 1 and noting that by definition the null space of Q_2^\top equals the column space of Φ , we have $x(0)$ is the unique solution.

Proof of C1 by contradiction: Suppose there exist $2s$ columns of Q_2^\top that are linearly dependent. Then, there exists $E_0 \neq 0$ such that $Q_2^\top E_0 = 0$ where $|\text{supp}(E_0)| \leq 2s$. Since the null space of Q_2^\top equals the column space of Φ , there exists z such that $E_0 = \Phi z$ (i.e., E_0 is in the column space of Φ). Then, $|\text{supp}(\Phi z)| = |\text{supp}(E_0)| \leq 2s$ (contradiction). ■

The sufficient condition, provided in Lemma 2, for the existence of a unique solution to (4.10) is hard to check as it requires satisfiability of the condition for all $z \in \mathbb{R}^n \setminus \{0\}$. In the following Theorem, we prove an equivalent, yet simple-to-check, sufficient condition that only needs to be verified for the eigenvectors of A .

Theorem 1 *Let $A \in \mathbb{R}^{n \times n}$, $C \in \mathbb{R}^{p \times n}$. Assume that C is full rank, (A, C) is observable and A has n distinct positive eigenvalues such that $0 < \lambda_1 < \lambda_2 < \dots < \lambda_n$. Define:*

- $s_i \triangleq |\text{supp}(Cv_i)|$, where v_i is an eigenvector of A ,
- $\mathcal{S} \triangleq \{s_1, s_2, \dots, s_n\}$,
- For every $m \in \{2, \dots, n\}$, let \mathcal{S}_m be any subset of \mathcal{S} with m elements, define $T_{\mathcal{S}_m} \triangleq \frac{(m-2) \cdot p + \min \mathcal{S}_m}{\max \mathcal{S}_m - 2q}$. Then T_m is chosen such that $T_m > T_{\mathcal{S}_m}$ for all subsets \mathcal{S}_m , i.e., all subsets of m elements from the set \mathcal{S} .

Choose T such that $T \geq \max\{T_2, \dots, T_n\}$. Then, the following are equivalent:

- (i) $\forall v_i \in \mathbb{R}^n$ where $Av_i = \lambda_i v_i$, $|\text{supp}(Cv_i)| > 2q$
- (ii) $\forall v_i \in \mathbb{R}^n$ where $Av_i = \lambda_i v_i$, $|\text{supp}(\Phi v_i)| > 2q \cdot T$
- (iii) $\forall z \in \mathbb{R}^n \setminus \{0\}$, $|\text{supp}(\Phi z)| > 2q \cdot T$

Theorem 1 states that if the feedback system and the secure estimator are designed such that all the conditions in the theorem are satisfied, then our proposed secure estimator can guarantee accurate correction of q errors by checking the following very simple condition:

$$\forall v_i \in \mathbb{R}^n \text{ where } Av_i = \lambda_i v_i, |\text{supp}(Cv_i)| > 2q. \quad (4.11)$$

4.3.2.1 Proof of Theorem 1

In order to prove Theorem 1, we make use of Lemma 4 and Proposition 2 (see Appendix).

First, it is simple to prove that (i) and (ii) are equivalent: $|\text{supp}(\Phi v_i)| = \sum_{k=0}^{T-1} |\text{supp}(CA^k v_i)| = \sum_{k=0}^{T-1} |\text{supp}(\lambda_i^k Cv_i)| = T \cdot |\text{supp}(Cv_i)|$.

Second, we want to show that (ii) and (iii) are equivalent. The direction (iii) \implies (ii) is trivial, since (ii) is a specific case of (iii) with $z = v_i$. The other direction is more complex. Note that A is diagonalizable, therefore its eigenvectors form a basis for \mathbb{R}^n . Now consider the decomposition of z in the eigenbasis of A , i.e. $z = \sum_{i=1}^n \alpha_i v_i$ with $\alpha_i \neq 0$ for at least one i .

1. $m = 1$: Suppose there exists $z \in \mathbb{R}^n \setminus \{0\}$ such that $|\text{supp}(\Phi z)| \leq 2q \cdot T$. Without loss of generality, let $\alpha_i \neq 0$ and $\alpha_j = 0$ for all $j \neq i$, then, $2q \cdot T \geq |\text{supp}(\Phi z)| = |\text{supp}(\alpha_i \Phi v_i)| = |\text{supp}(\Phi v_i)|$ for all T (contradiction, $\because \forall v_i, |\text{supp}(\Phi v_i)| > 2q \cdot T$).
2. $m = 2$: By Lemma 4, if we choose $T > T_2$.
3. $m \geq 3$: By Lemma 4 and Proposition 2, if we choose $T > T_m$ for each value of m , respectively.

We need to choose T to satisfy the worst case for any m such that $n \geq m \geq 2$. Thus, if $T \geq \max\{T_2, \dots, T_n\}$, then (ii) and (iii) are also equivalent. Note that T_m is chosen to satisfy $T_m > T_{\mathcal{S}_m}$ for all subsets \mathcal{S}_m , i.e., $T_m = \max_{\mathcal{S}_m} (T_{\mathcal{S}_m}) + 1$. \blacksquare

4.3.3 Number of Correctable Errors

Given that the set of attacked nodes can change over time and $e(k)$ satisfies $|\text{supp}(e(k))| \leq q$ for all k , we prove in Proposition 1 (see below) that the maximum number of correctable errors (as defined in Definition 4.3.1) by our decoder is $\lceil p/2 - 1 \rceil$, where p is the number of measurements.

Proposition 1 *Let $A_0 \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$ and $C \in \mathbb{R}^{p \times n}$ and assume that the pair (A_0, B) is controllable, C is full rank and each row of C is not identically zero. Then there exists a finite set $F \subset \mathbb{R}_+$ such that for any choice of n numbers $\lambda_1, \dots, \lambda_n \in \mathbb{R}_+ \setminus F$ such that $0 < \lambda_1 < \dots < \lambda_n$, there exists $G \in \mathbb{R}^{m \times n}$ such that:*

- The eigenvalues of the closed-loop matrix $A (= A_0 + BG)$ are $\lambda_1, \dots, \lambda_n$.
- If the pair (A, C) is observable, then the number of correctable errors for the pair (A, C) is maximal after $T = \max\{n, T^*\}$ time steps and is equal to $\lceil p/2 - 1 \rceil$, where T^* is the value of T from Theorem 1.

Proof: The proof for Proposition 4 in [24] shows that if the chosen poles $\lambda_1, \dots, \lambda_n$ are distinct, positive and do not fall in some finite set F , then there is a choice of G such that the eigenvalues of $A (= A_0 + B)$ are exactly $\lambda_1, \dots, \lambda_n$, and the corresponding eigenvectors v_i are such that $|\text{supp}(Cv_i)| = p$. Thus, by Theorem 1, the number of correctable errors for (A, C) is $\lceil p/2 - 1 \rceil$. ■

In addition, recall that $E_{q,T}$ consists of the error vectors $e(0), \dots, e(T-1)$ stacked vertically and our proofs for the existence of a unique solution to (4.10) are independent of how the individual error (nonzero) terms are distributed in the vector $E_{q,T}$. Thus, we can remove the assumption: $|\text{supp}(e(k))| \leq q$ for all k , and allow $e(k)$ to appear in an arbitrary fashion, e.g. $|\text{supp}(e(0))| = 2q$ and $|\text{supp}(e(1))| = 0$, as long as $\sum_{k=0}^{T-1} |\text{supp}(e(k))| \leq q \cdot T$, then our q -error-correcting estimator can still recover the true states. In other words, our proposed secure estimator can protect the system against more general attacks where the number of attacked sensors is not necessarily less than or equal to q at every time step.

4.4 Estimator Design

In the classical error correction problem, to ensure accurate estimation, the coding matrix must satisfy the Restricted Isometry Properties (RIP) conditions [10], which are extremely difficult to check in general. In practice, Theorem 1.4 from [10] is almost always used to design a coding matrix *a priori*. This theorem states that a coding matrix whose entries are sampled from independent and identical distributions satisfies the RIP condition with overwhelming probability. In secure estimation, however, it is impossible to choose such a coding matrix *a priori* because it is the observability matrix Φ , which is structurally constrained: as shown in (4.5), Φ consists of CA^k 's where $k = \{0, \dots, T-1\}$. In this Section, we use Condition 4.11 from Theorem 1, the results from Proposition 4 in [24] and state feedback to design a matrix Φ for accurate estimation.

4.4.1 System and Estimator Design

Since conditions (i) and (iii) in Theorem 1 are equivalent, condition (i) can be used to design a state feedback controller such that the closed system can achieve accurate estimation. Therefore, given a controllable open-loop pair (A_0, B) , design C

and choose an adequate feedback control law $u(k) = Gx(k)$ and construct a secure estimator such that:

1. Each row of C is not identically zero, and C is full rank;
2. The closed-loop matrix A ($= A_0 + BG$) has n distinct positive eigenvalues: $0 < \lambda_1 < \lambda_2 < \dots < \lambda_n$;
3. (A, C) is observable;
4. The length of the sliding window of measurements T of the estimator satisfies Theorem 1²;
5. Maximize q subject to: $\forall v_i \in \mathbb{R}^n$ where $Av_i = \lambda v_i$, $|\text{supp}(Cv_i)| > 2q$.

Without loss of generality, the first condition holds. For example, if there exists a zero row in C , we can simply remove that row from C without changing the system's behavior. Conditions 2, 3 and 4 are required for equivalence in Theorem 1. The last condition is needed for accurate estimation and for maximizing the number of correctable attacks. From Proposition 1 the maximum number of correctable attacks can be achieved when $|\text{supp}(Cv_i)| = p$ (i.e., the number of measurements) for all eigenvectors of A .

Conditions 2, 3 and 5 depend on the feedback controller. So how do we choose a controller that achieves good performance in both control and secure estimation? Below, we describe an approach that we have taken in all simulations in this chapter, and has proved to work well. This is by no means the only method. First, we design a controller that achieves good control, for example, Linear Quadratic Regulators (LQR), which are optimal with respect to a certain quadratic cost function. However, these controllers may not have good secure estimation properties, meaning the value of q that satisfies $|\text{supp}(Cv_i)| > q$ for all eigenvectors of A may be small, i.e., the resulting estimator can only correct few attacks. It is often easy to increase the value of q and make the estimator more resilient to attacks by slightly perturbing the closed-loop poles from those resulting from the LQR controller, such as placing the poles closer to the origin, and making the poles more spread out amongst themselves. We chose to keep the perturbations small as to not lose too much control performance. Although this is a heuristic method, it is relatively easy to carry out in order to satisfy the above conditions; whereas in the classical error correction method [10], checking whether a coding matrix satisfies RIP is extremely difficult.

To summarize, we start from some optimal controller which may not result in a good estimator, then we perturb the closed-loop poles slightly to improve the resulting

²We found that much smaller T 's are often sufficient for good secure estimation performance, i.e., to perfectly recover the attack signals. In all simulations in this chapter, $T = n$ is used, where n is the number of states.

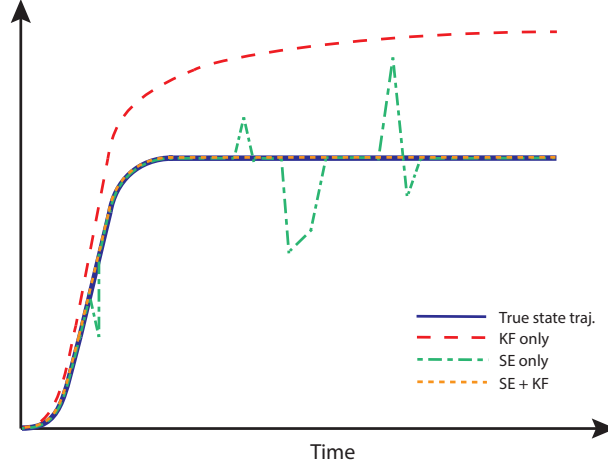


Figure 4.1: Illustrative comparison of three schemes: KF only (KF), secure estimator only (SE), secure estimator with a KF (SE+KF). KF fails to estimate the true state as attack signal is non-Gaussian. SE correctly estimates the system state most of the time but has occasional large estimation attacks. SE+KF tracks the true state trajectory perfectly.

estimator's secure estimation capability. Therefore there is a trade-off between a system's control and secure estimation performances, and the feedback controller can be designed to achieve a desired trade-off between them.

4.4.2 Combination of Secure Estimation and Kalman Filter

Consider the state estimation problem for the following LTI system under attack:

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) \\ y(k) &= Cx(k) + e(k) + v(k), \end{aligned} \quad (4.12)$$

where x , y , u and e are as defined in (4.3); and v is a zero mean independent and identically distributed (i.i.d.) Gaussian measurement noise.

A KF can be used to estimate the states by modeling the attack signal as noise. More specifically, define a new measurement noise $\bar{v}(k) = e(k) + v(k)$ to give a new measurement equation $y(k) = Cx(k) + \bar{v}(k)$. A KF can then estimate the states from the inputs $u(k)$ and the corrupted measurements $y(k)$ [53]. One caveat with this method is that KFs assume zero mean and i.i.d. white Gaussian measurement noise, however, attack signals are usually erratic and may be poorly modeled by Gaussian processes [53], i.e., $e(k)$ and consequently, $\bar{v}(k)$ may not be Gaussian. Take GPS spoofing attacks for example, attack signals are often structured to resemble normal GPS signals or can be genuine GPS signals captured elsewhere. When the system is subjected to attacks that are poorly modeled by Gaussian processes, it is reasonable

to expect KFs to fail to recover the true states. Figure 4.1 gives an illustrative example where an attack signal that increases linearly with time is injected into the measurements of state x_i . The red dashed line shows a plausible estimated state trajectory from a KF.

On the other hand, our proposed secure estimator does not assume the attack signal to follow any model, and therefore, it works for arbitrary and unbounded attacks. The only assumption is that the number of attacked sensors is sparse, i.e., less than $\lceil p/2 - 1 \rceil$. As the set of attacked sensors becomes less sparse, our secure estimator occasionally fails to recover the true states. The green dashed line in Figure 4.1 depicts a possible result from this estimator: the estimated state trajectory follows the true trajectory most of the time with occasional attacks. Based on these observations, we propose to combine our secure estimator with a KF to improve its practical performance, as detailed in Algorithm 1.

Algorithm 2 Combined secure estimator with KF

```

1: Initialize the KF
2: for each  $k$  do
3:   if  $k \geq T$  then
4:     Estimate the attack signal at time  $k$ ,  $\hat{e}(k)$ , using secure estimator
5:   else
6:     Set  $\hat{e}(k) = 0$ 
7:   end if
8:   Form a new measurement equation:  $\tilde{y}(k) = Cx(k) + \tilde{v}(k)$ , where  $\tilde{y}(k) =$ 
      $y(k) - \hat{e}(k)$  and  $\tilde{v}(k) = e(k) - \hat{e}(k) + v(k)$ 
9:   Apply standard KF using  $u$  and  $\tilde{y}$ 
10: end for

```

The intuition is that the secure estimator acts as a pre-filter for the KF, so that $\tilde{v}(k)$ is close to a zero mean i.i.d. Gaussian process even when the true attack signal $e(k)$ is not. More specifically, the secure estimator usually perfectly recovers $e(k)$, thus $e(k) - \hat{e}(k) = 0$ and $\tilde{v}(k) = v(k)$. What happens when the secure estimator fails? Equation (4.5) shows that the estimated state at time k , $\hat{x}(k)$, does not directly depend on the estimated state at another time point $\hat{x}(\tau)$ ($t \neq \tau$). As a result, when the secure estimator fails, its estimation error, $e(k) - \hat{e}(k)$, appears to be quite random. Putting these together: $\tilde{v}(k) = e(k) - \hat{e}(k) + v(k)$ is closer to a zero mean i.i.d. white Gaussian process than $\bar{v}(k)$ (i.e., the corresponding measurement noise if a KF is applied directly to estimate the states), which improves the KF's performance. Finally, the *if* statement in Algorithm 1 ensures that the secure estimator always has access to T past measurements, as required by Theorem 1.

Next, we demonstrate the effectiveness of our proposed method through simulations of a UAV under two types of adversarial attacks, which also provides a realistic example illustrating the behaviors described in this section.

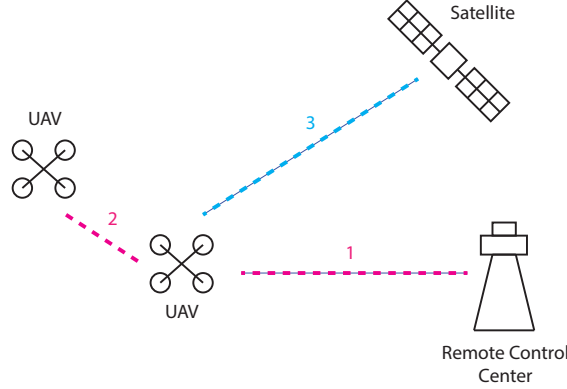


Figure 4.2: Different communication channels that are subject to adversarial attacks: Channels 1 and 2 are vulnerable to attack in the MITM attack example (Section 5.3.1) and Channel 3 is vulnerable to attack in the GPS spoofing example (Section 5.3.2).

4.5 Numerical Examples

On February 15, 2015, the Federal Aviation Administration proposed to allow routine use of certain small, non-recreational UAVs in today’s aviation system [19]. Thus in the near future, we may see thousands of UAVs such as Amazon Prime Air [1] and Google Project Wing vehicles [29] sharing the airspace simultaneously. To ensure safety of this immense UAV traffic, UAVs may periodically update their position and velocity measurements wirelessly to a RCC for traffic management (Channel 1 in Figure 4.2). At the same time, UAVs may broadcast this information to other UAVs in its vicinity for collaborative collision avoidance (Channel 2 in Figure 4.2). Finally, autonomous UAVs may use GPS for their position measurements (Channel 3 in Figure 4.2). All these communication channels are subject to cyber attacks. If corrupted information are used in collision avoidance or path planning algorithms, they can lead to possible collisions or loss of UAVs, causing physical and financial damage and even injury to civilians. To help protect against these attacks and consequences, participating entities such as the UAVs and the RCC can use secure estimation to estimate a target UAV’s true position and velocity before using any received information for collision avoidance, for instance. In this section, we focus on 2 types of adversarial cyber attacks on UAVs and demonstrate the effectiveness of our secure estimator through simulations.

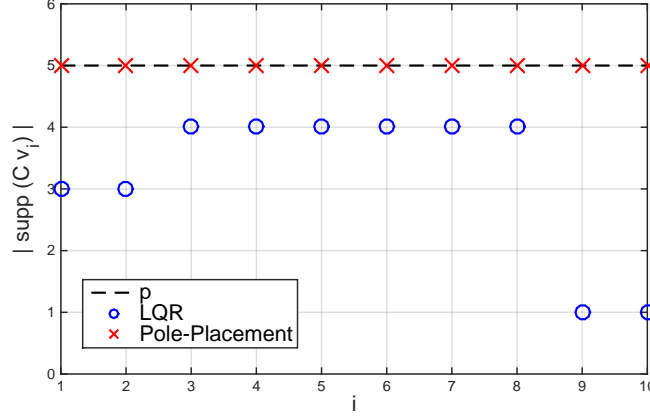


Figure 4.3: $|\text{supp}(Cv_i)|$ for all eigenvectors v_i of closed-loop matrix A for 2 feedback controllers: a LQR and a controller designed by pole-placement. Black dashed line is at $p = 5$, i.e., the number of measurements.

4.5.1 UAV Model

We consider a quadrotor with the following dynamics:

$$\begin{aligned} x(k+1) &= A_0x(k) + Bu(k) + k + w(k) \\ y(k) &= Cx(k) + e(k) + v(k) \end{aligned} \quad (4.13)$$

where $x = [p_x, v_x, \theta_x, \dot{\theta}_x, p_y, v_y, \theta_y, \dot{\theta}_y, p_z, v_z]^T$ is the state vector. p_x , p_y and p_z represent the quadrotor's position along the x , y and z axis, respectively. v_x , v_y and v_z represent its velocities. θ_x and θ_y are the pitch and roll angles respectively, $\dot{\theta}_x$ and $\dot{\theta}_y$ are their corresponding angular velocities. $u = [\theta_{r,x}, \theta_{r,y}, F]^T$ is the input vector: $\theta_{r,i}$ is the reference pitch or roll angle, and F is the commanded thrust in the vertical direction. $y = [\tilde{p}_x, \tilde{p}_y, \tilde{p}_z]^T$ represents compromised position measurements from the GPS under attack signal e . w and v represent process and measurement noise respectively. k is a constant vector which represents gravitational effects, and can be dropped without loss of generality because we can always subtract it out in u . $A_\theta^{i,j}$ refers to the ij -th entry of the subsystem matrix of the discretized rotational dynamics A_θ , and B_θ^i refers to the i -th entry of the input-to-state map B_θ for the discretized rotational dynamics. T_s is the discrete time step, g is the gravitational acceleration, m is the mass of the quadrotor and K_T is a thrust coefficient. Further details about this model and its derivation can be found in [7]. Finally, the matrix C depends on the particular measurements taken in each example.

4.5.2 Estimator Design via Pole-Placement

Assume that the UAV uses the state feedback control law $u(k) = Gx(k)$, where G is the feedback matrix which can be designed³. If the pair (A_0, B) is controllable, then we can choose G to place the closed loop poles anywhere in the complex plane. We first design a Linear Quadratic Regulator (LQR) and evaluate its secure estimation performance by checking whether the sufficient condition for q -attack correction (i.e., $|\text{supp}(Cv_i)| > q$ for all i) holds. Figure 4.3 shows the results for a matrix $C \in \mathbb{R}^{5 \times 10}$ (i.e., 5 measurements) and observe that $|\text{supp}(Cv_i)| < p = 5$ for $i = 1, 2, 9$ and 10 . Furthermore, $|\text{supp}(Cv_i)| = 1 > 0$ for $i = 9$ and 10 , therefore the resulting secure estimator can correct zero attacks! To improve the secure estimation performance, we perturb the closed-loop poles slightly until $|\text{supp}(Cv_i)| = p$ for all i , as shown in Figure 4.3. Therefore the resulting secure estimator can achieve the maximum number of correctable attacks within the limits of p (i.e., the number of measurements). By keeping the perturbations on the poles small, our final controller achieves both good control and estimation performances (see Figure 4.4).

4.5.3 UAV under Adversarial Attack

4.5.3.1 Man-In-The-Middle (MITM) Attack in Communication with a RCC or with other UAVs

In this section, we consider MITM attacks targeted at Channels 1 and 2 in Figure 4.2, where a malicious agent spoofs the information being sent and/or received over these channels. The goal of the RCC or other UAVs is to accurately estimate the true flight path of a target UAV from corrupted measurements. Note that the true path of the target UAV is unaffected by the attack. Assume that the attacker spoofs the position measurements in order to deceive the receiver that the target UAV is deviating in the x -direction, i.e., she injects a continuous and increasing signal in the x -position measurement. To make the estimation task even harder for the receiver, the attacker also injects a random Gaussian noise to an additional measurement, and the choice of this measurement can change at each time step.

In this example, we first demonstrate the effectiveness of our proposed estimator design via pole-placement method by comparing the estimation performance of the estimator resulting from (1) a LQR controller and (2) a controller designed using pole-placement as described in the previous section. Throughout this example, $y \in \mathbb{R}^5$, measurements include the x , y and z positions and 2 additional randomly selected states.

The left plots in Figure 4.4 show the true attack signal on all 5 sensors (solid lines) and the estimated attack signals (dashed lines) by the secure estimator if the feedback

³In the GPS spoofing example, direct uncorrupted state measurements are not available. Therefore a KF is used to give estimated states which are then used for state feedback control.

controller is a LQ regulator (top) or one designed via pole-placement (bottom). It is obvious that the latter estimates the attack signal much more accurately. The right plots of this figure highlights this observation by explicitly showing the estimation error of the attack signal for each measurement.

The same information is shown in Figure 4.5, where each row corresponds to one sensor, and the first 3 rows are the x , y and z position measurements, respectively. This figure highlights three points: first, the attacked sensors change with time; second, the number of attacked sensors at each time k is less or equal to 2; third, only position measurements are corrupted.

Note that the poor performance is not an inherent feature of LQR. Since LQR does not consider the conditions for secure estimation, there is no guarantee of good secure estimation performance. On the other hand, if we satisfy the guaranteed conditions for accurate secure estimation, then we can correctly estimate the true attack signals, i.e., achieve good secure estimation performance (although we may lose some control performance).

Next, we implement feedback controller (2), i.e., one designed with pole-placement, and compare the performance of three different state estimation schemes: (a) KF only (KF), (b) secure estimator only (SE), and (c) secure estimator combined with KF (KF+SE). Figure 4.6 shows the estimated flight paths by all three methods. The true path of the UAV (solid blue line) starts from the position marked by the blue triangle and ends at the position marked by the blue square. KF fails to filter out the attack signal in the x -position measurements as the attack is highly non-Gaussian, and the estimated trajectory (dashed red line) significantly differs from the true one. On the other hand, SE correctly estimates some portions of the trajectory and the final position of the vehicle, nevertheless it produces spontaneous attacks in the x direction. Finally the combined method KF+SE perfectly recovers the true path of the target UAV.

4.5.3.2 GPS Spoofing

In this section, we focus on adversarial attacks in the GPS navigation system (Channel 3 in Figure 4.2). Consider the scenario where a UAV uses a Linear Quadratic Gaussian (LQG) controller to follow a desired path, $x_r(k)$, designed by LQ control. In other words, a KF takes compromised and noisy measurements $y(k)$ and outputs a state estimate $\hat{x}(k)$, which is then used for state feedback control: $u(k) = G(\hat{x}(k) - x_r(k))$, where G is the feedback matrix. Note that in the previous example (Section 4.5.3.1), the feedback controller had access to uncorrupted state measurements $x(k)$, therefore the true path of the UAV is unaffected by attacks. On the other hand, in this example, the UAV uses estimated states $\hat{x}(k)$ for feedback control and path following. Hence, if measurements are corrupted and the state estimates are poor, then the UAV may not be able to follow its desired path and may deviate away from it. The goal is to correctly estimate the true states of the UAV

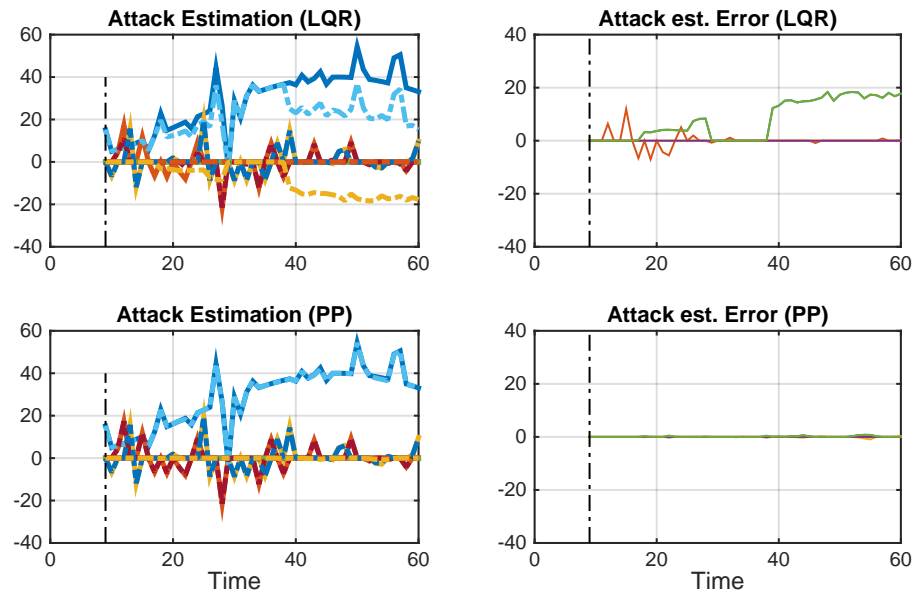


Figure 4.4: True attack signal, estimated attack signal and estimation error in the attack signal of the estimator (SE) with 2 different feedback controllers: LQR, controller designed via pole-placement (PP); with 5 measurements. In the left plots, solid lines are true attack signals, dashed lines are estimated signals. The right plots show the estimation error in the attack signal.

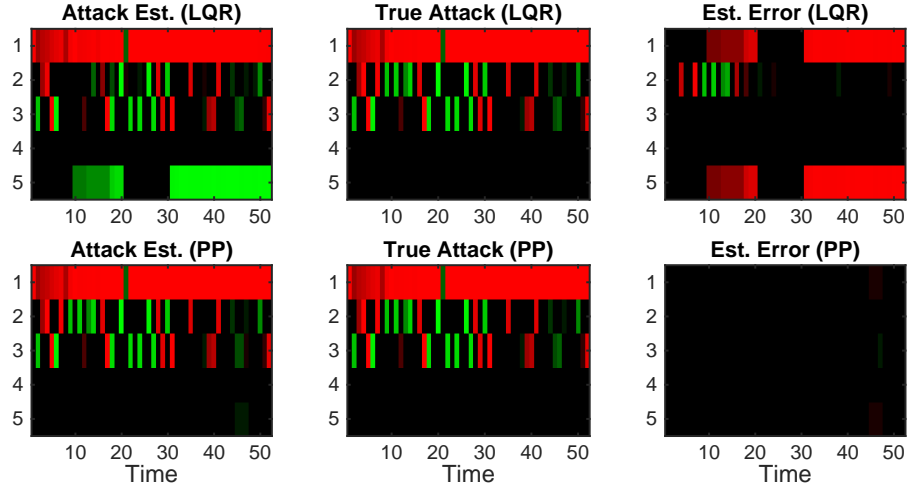


Figure 4.5: Estimated attack signal, true attack signal and estimation error in the attack signal of the estimator (SE) with 2 different feedback controllers: LQR, controller designed via pole-placement (PP); with 5 measurements. Left column shows estimated attack signals. Middle column shows true attack signal. Right column shows estimation error. Each row corresponds to one type of measurement. Red pixels indicate positive values, green pixels are negative values and black indicates zero.

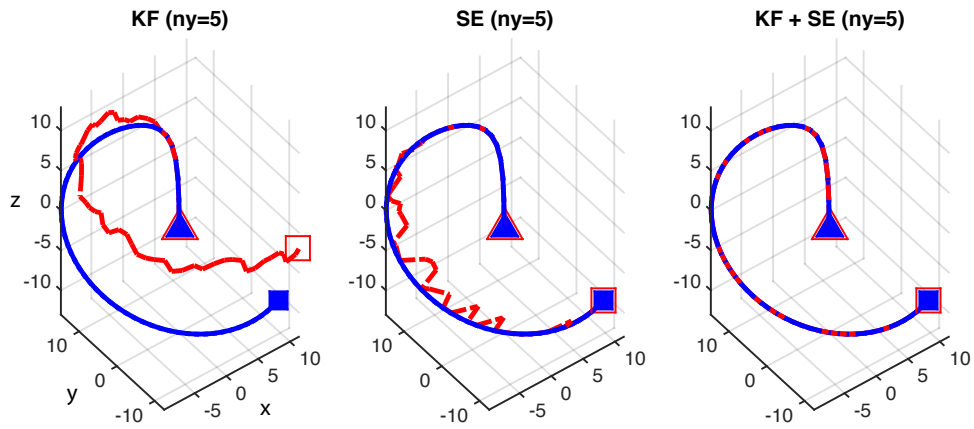


Figure 4.6: Estimated UAV trajectory by three methods under MITM attack: KF only (KF), secure estimator only (SE), secure estimator with KF (KF+SE). Solid blue lines are the true UAV trajectories. They start from the blue triangle and end at the blue square. Red dotted lines represent estimated trajectories by each method, with 5 measurements.

and therefore, follow the desired path. Assume an attacker spoofs the GPS position measurements in order to deviate the UAV from its planned path. She injects a sinusoidal signal to x -position measurement, as well as a Gaussian noise to a randomly chosen position measurement at each time step.

In this example, we explore the effect of the number of sensor measurements on the secure estimation performance of two schemes: (a) KF only, (b) KF+SE. We first assume that the UAV only uses GPS for navigation, i.e., 3 positional measurements. Figure 4.7 shows that KF completely fails to estimate the attack signal (KF, $n_y = 3$, plots in Row 1), consequently the actual UAV trajectory (red dashed line) deviates significantly from its desired path (solid blue line) as shown in Figure 4.8, and deviations are largest along the x - and z -axis (Figure 4.9). On the other hand, Figure 4.7 (KF + SE, $n_y = 3$) shows that KF+SE's estimated attack signals are significantly more accurate with only a small estimation error in the x -position (plots in Row 2). Therefore the UAV can follow its planned path much more closely (Figures 4.8 and 4.9). Recall from Proposition 1 that the maximum number of correctable attacks for a system with p measurements is $\lceil p/2 - 1 \rceil$, which equals 1 in this case. There are at most 2 attacked sensors at any time k in this example, which exceeds the above limit. This explains the estimation error in the x -position. Despite this small estimation error, the combined scheme KF+SE still outperforms the KF on its own.

We now show the effect of increasing the number of measurements (n_y , or equivalently p) through sensor fusion, on the estimation performance and consequently, the UAV's path following performance. Autonomous UAVs often use IMUs in addition to GPS for navigation, the former provides additional measurements such as the UAV's velocities, pitch and roll angles. Figure 4.7 shows that increasing the number of measurements has no effect on the KF's estimation accuracy (compare plots in Rows 1, 3 and 5). Even when 8 measurements are used the UAV equipped with a KF still fails to follow the desired path (Figures 4.8 and 4.9). On the other hand, increasing the number of measurements improves the estimation performance of the secure estimator (SE) and consequently the performance of the combined scheme KF+SE (compare plots in Rows 2, 4 and 6 in Figure 4.7). Observe that for both 3 and 5 measurements, the combined scheme KF+SE perfectly estimates the attack signals and therefore, can completely subtract them out from the corrupted measurements. As a result, the UAV can follow its original planned path perfectly (KF + SE $n_y = 5$, KF + SE $n_y = 8$ in Figures 4.8 and 4.9).

4.6 Conclusion

In this chapter, we consider the problem of secure estimation for CPS under adversarial attacks. Unlike [24] where the attacked sensors are assumed to be fixed, we allow the set of attacked sensors to change over time, and propose a computationally efficient secure estimator for the latter scenario that works for arbitrary

and unbounded attacks. In addition, we propose to combine the secure estimator with a KF for improved practical performance. We demonstrate through numerical examples, that our proposed secure estimator based KF outperforms standard KF. Furthermore, we illustrate practical applications of secure estimation in UAVs under adversarial cyber attacks. This is important not only for today's aviation system but also UAV delivery systems in the near future.

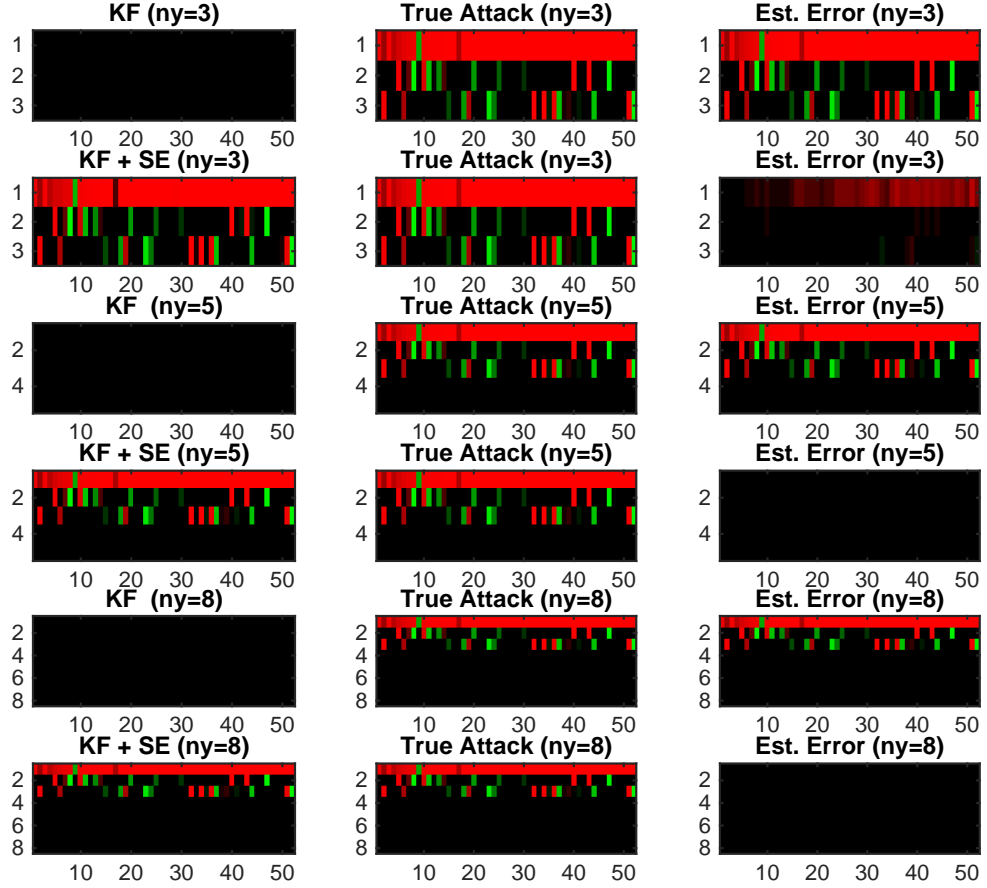


Figure 4.7: Estimated attack signal, true attack signal and estimation error in different cases: KF and KF+SE, each using 3, 5 and 8 different measurements. Left column shows estimated attack signals. Middle column shows true attack signal. Right column shows estimation error. Each row corresponds to one sensor measurement and the first three rows in each plot are the x , y and z position measurements, respectively. Red pixels indicate positive values, green pixels are negative values and black indicates zero.

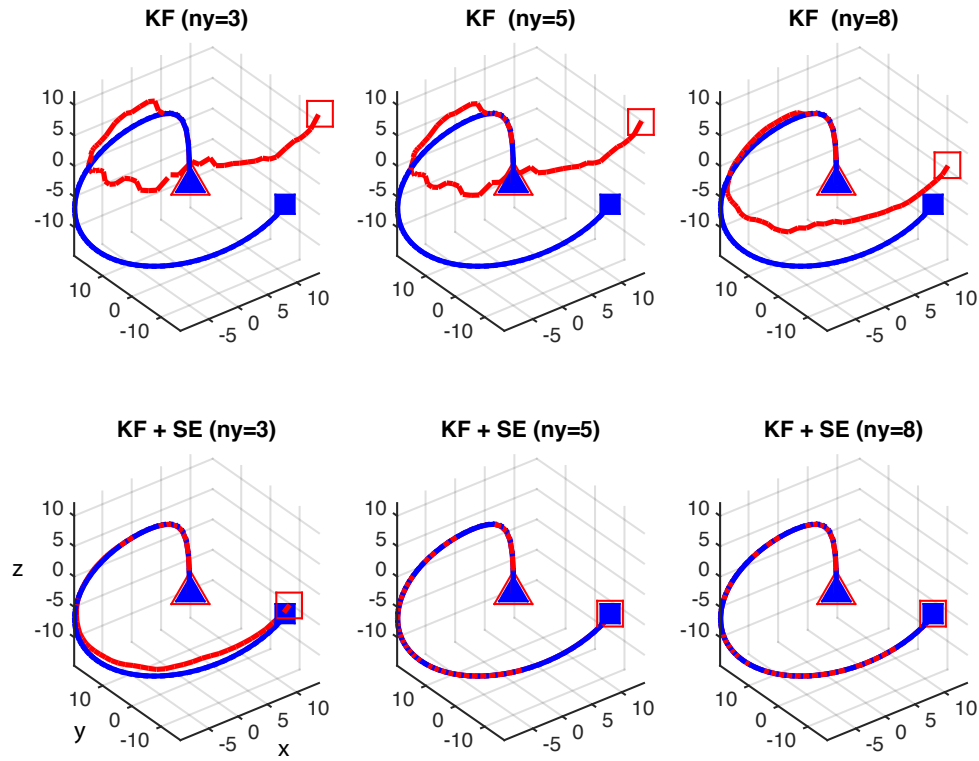


Figure 4.8: Desired and actual UAV trajectory in different cases: KF and KF+SE, each using 3, 5 and 8 different measurements. Blue solid lines are the desired trajectory. Red dash lines are the actual UAV trajectory under adversarial attack.

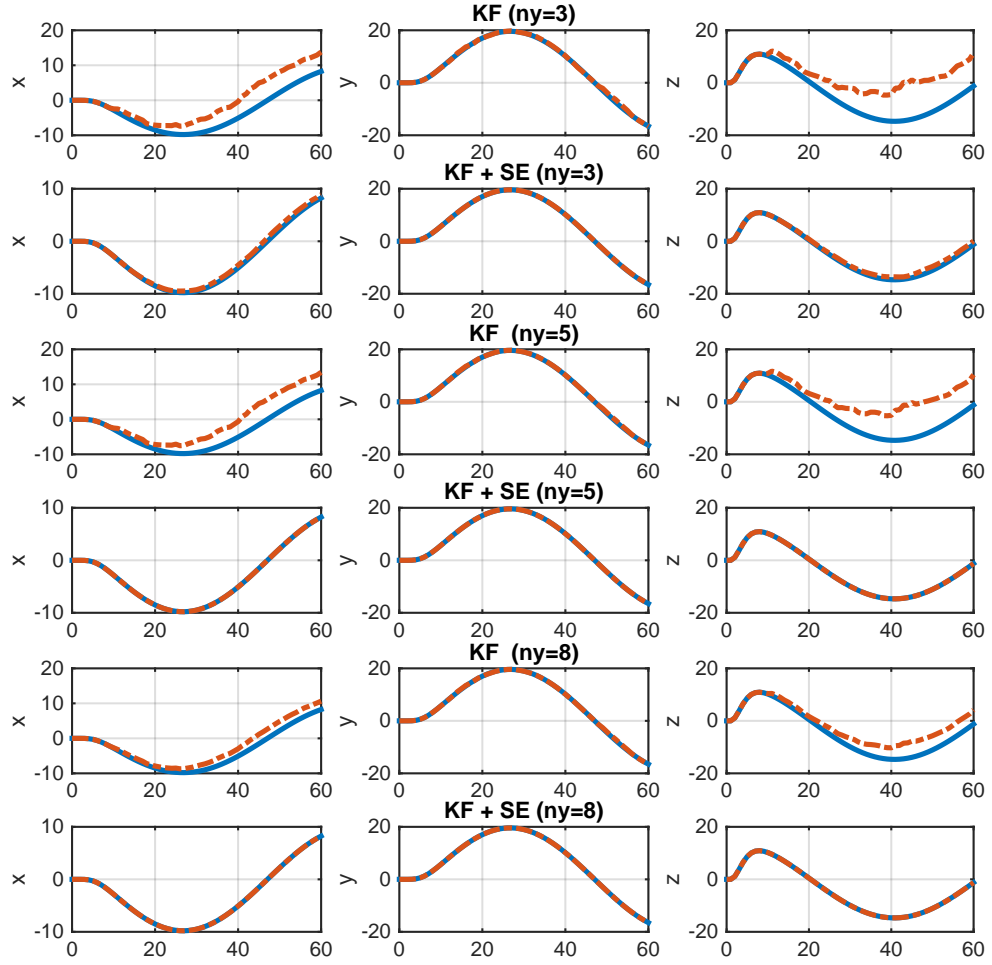


Figure 4.9: Desired and actual UAV trajectory in different cases: KF and KF+SE, each using 3, 5 and 8 different measurements. Blue solid lines are the desired x , y and z trajectories. Red dashed lines are the actual UAV trajectories under adversarial attack.

Chapter 5

Secure Estimation for Nonlinear Power Systems

Many approaches in the literature that address the secure state estimation problem are based on linear dynamical systems. Hence, the existing secure estimators can be applied to nonlinear dynamical systems if we linearize these systems. It is well known that the linearization of nonlinear dynamical systems can result in the following drawbacks:

1. Linearization is reliable if the higher order terms in the Taylor series expansion can be eliminated; otherwise, the linearized model may perform poorly.
2. Linearization can be applied when all the eigenvalues of the Jacobian matrix have nonzero real part. However, this is not always the case.

For example, linearized power system models are only valid under small perturbations in the system at hand. Under a severe disturbance, such as a single or multi-phase short-circuit or a generator loss, the linearized model does not remain valid [51], [95]. Therefore, the existing techniques lack performance guarantees when the system undergoes large perturbations which are typical of highly loaded practical systems. To overcome the above drawbacks, we develop a secure state estimation method without linearization or calculation of Jacobian matrices. Note that feedback linearization techniques transform the nonlinear system into an equivalent linear system through a change of variables and a suitable control input. Even with such techniques, the secure estimation problem for nonlinear dynamical systems is a nonlinear problem.

In this chapter, we investigate the secure estimation of the state of a nonlinear dynamical system from a set of corrupted measurements for two classes of nonlinear systems, and propose a technique which enables us to perform secure state estimation for those systems. We then illustrate how the proposed nonlinear secure state estimation technique can be used to perform estimation in the cyber layer of interconnected power systems under cyber-physical attacks and communication failures.

In particular, we focus on an interconnected power system comprising several synchronous generators, transmission lines, loads, and energy storage units, and propose a secure estimator that allows us to securely estimate the dynamic states of the power network. Finally, we numerically demonstrate the effectiveness of the proposed secure estimation algorithm, and show that the algorithm enables the cyber layer to accurately reconstruct the attack signals.

This chapter is an adaptation of the paper in [44].

5.1 Introduction

To overcome the limitations of applying linear system based secure state estimation methods on nonlinear systems, we investigate the secure estimation of the state of a nonlinear dynamical system from a set of corrupted measurements. As in Chapter 4, we do not make any assumption on the sensor attacks or corruptions (i.e., corruptions can follow any particular model). Our only assumption concerning the corrupted sensors is about the number of sensors that are corrupted due to attacks or failures. We consider two classes of nonlinear systems, and design secure state estimators for these assuming that the set of attacked sensors can change with time. A practical example of such a cyber attack is described in [56], where a multi-switch attack, in which different switches in a power network are attacked at different times, is designed to lead to stealthy and wide-scale cascading failures in the power system. We then propose a technique which enables us to transform the nonlinear dynamics into a set of linear equations, and apply the classical error correction method to the equivalent linear system. The proposed secure state estimators are computationally efficient and can be solved exactly without iteration. In addition, our estimator relies on the observability of the transformed linear system, which is much simpler to check than verifying the observability of nonlinear systems.

The work closest to ours is [84] in which Shoukry *et al.* studied differentially flat nonlinear systems under sensor attack and assumed that the set of attacked sensors do not change with time. Using s -sparse observability for nonlinear systems, the authors proposed a combinatorial estimator, and an iterative satisfiability modulo theory-based algorithm to solve the resulting combinatorial estimation problem. However, it may be hard to check the observability of nonlinear systems, and the assumption of fixed attacked nodes may be restrictive.

To illustrate how our proposed secure state estimator approach can be applied to practical systems, we focus on an interconnected power system comprising several synchronous generators, transmission lines, buses, and energy storage units. We assume that all the physical devices are controlled via a WACS as well as local controllers, and that these control systems use the synchrophasor technology, PMUs, to

maintain the system's stability¹. The WACS and local controllers employ advanced data acquisition, communications, and control to enable increased efficiency and reliability of power delivery [12], [58], [66], [96]. Several methods for power system state estimation have been proposed [26], [6], [98], [79], [46], [45]. All these methods rely on the linearization. To overcome the drawbacks of linearization, Wang *et al.* [95] develop a dynamic state estimation method that requires neither linearization nor calculation of Jacobian matrices. However, the authors only consider Gaussian noise. Extensive work has been done on monitoring and autonomous feedback control for WACS [96], and on secure state estimation of static states [41], [50]. However, these works have not studied how to identify cyber-physical attacks or communication failures when dynamic states such as generator' phase angles are estimated, and how to perform secure state estimation (dynamic state estimation) for the WACS.

We focus on secure estimation for the wide area control system of the power network assuming that the installed PMUs at different generator buses are connected through a communication network which sends PMU measurements to the WACS as well as the local controllers in the power network. We assume that the communication channels from the WACS to the generators are secured while other channels and PMUs are not secured and are subject to cyber attacks and failures. Therefore, the WACS needs to perform secure state estimation to reconstruct the system's states before using the received data for computing wide area control signals, and to monitor the operation of local controllers. By using the developed secure estimation technique, we propose a secure state estimator for the wide area control of the power system, and numerically show that the proposed algorithm significantly improves the performance of the cyber layer in power systems.

The chapter is organized as follows: In Section 5.2, we formulate the nonlinear state estimation problem and propose a solution technique for two classes of nonlinear systems. We then illustrate how the proposed secure state estimation approach can be applied to power systems in Section 5.3 and 5.4. Finally, in Section 5.5, we numerically demonstrate the effectiveness of the proposed secure estimation algorithm.

5.2 Secure Estimation for Nonlinear Systems

Consider a nonlinear dynamical system given by

$$\begin{aligned} x(k+1) &= Ax(k) + f(x(k), e(k)) + u(k) \\ y(k) &= Cx(k) + e(k) \end{aligned} \tag{5.1}$$

¹The secondary generation control in power systems is an example of such cyber-physical structures. In this system, measurements and control signals are telemetered to and from the generating units and that control center adjusts the set-point of each generator based upon the integral of the frequency error.

where $x(k) \in \mathbb{R}^n$ represents the state at time $k \in \mathbb{N}$, $A \in \mathbb{R}^{n \times n}$, $f(x(k), e(k)) : \mathbb{R}^n \times \mathbb{R}^p \rightarrow \mathbb{R}^n$ represent the system's dynamics, and $u(k) \in \mathbb{R}^n$ is a control input. $C \in \mathbb{R}^{p \times n}$ is the sensors' measurement matrix, $y(k) \in \mathbb{R}^p$ are the corrupted measurements at time $k \in \mathbb{N}$, and $e(k) \in \mathbb{R}^p$ represents attack signals injected by malicious agents at the sensors. In general, at each time instant, the system dynamics can be a function of the received measurements $y(k)$ as well as the state of the system $x(k)$. Since $y(k)$ can be expressed as a function of $x(k)$ and $e(k)$ using the measurement equation, we consider $f(x(k), e(k))$ to be a function of both $x(k)$ and $e(k)$.

Our goal is to reconstruct $x(k)$ in (5.1) by using the received measurements. Here, we do not assume the errors $e(k)$ follow any particular model. More precisely, the i -th element of $e(k)$ can take any value in \mathbb{R} . However, if sensor $i \in \{1, 2, \dots, p\}$ is not attacked, then necessarily the i -th element of $e(k)$ is zero. The only assumption concerning the corrupted sensors is the number of sensors that are attacked or corrupted due to failures. Our analytical results characterize the number of errors that can be corrected by a decoder.

Next, we focus on the problem of reconstructing state $x(k)$ for two classes of nonlinear systems.

5.2.1 Existence of Mapping Function with Error Correction

Let us assume that there exists a mapping function $g(y(k)) : \mathbb{R}^p \rightarrow \mathbb{R}^n$ such that

$$g(y(k)) = f(x(k), e(k)) \quad (5.2)$$

The mapping function $g(y(k))$ enables us to transform the nonlinear system in (5.1) into a linear system for which the error correction technique introduced in Section 4.3 can be used to reconstruct the initial state $x(0)$. To do so, we first use (5.1) and (5.2) to obtain $g(y(k)) = x(k+1) - Ax(k) - u(k)$ for all k . We then construct a vector Y as follows:

$$\begin{aligned} Y &= \begin{bmatrix} y(0) \\ y(1) - C(g(y(0)) + u(0)) \\ y(2) - C(Ag(y(0)) + Au(0) + g(y(1)) + u(1)) \\ \vdots \\ y(T-1) - C(A^{T-2}g(y(0)) + A^{T-2}u(0) + \dots) \end{bmatrix} \\ &= \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{T-1} \end{bmatrix} x(0) + \begin{bmatrix} e(0) \\ e(1) \\ e(2) \\ \vdots \\ e(T-1) \end{bmatrix} = \Phi x(0) + E \end{aligned} \quad (5.3)$$

where $E = [e(0); e(1); \dots; e(T-1)] \in \mathbb{R}^{pT}$ is the set of error vectors, and $\Phi = [C; CA; CA^2; \dots; CA^{T-1}]$.

We can now apply the error correction technique from Section 4.3 to the linear system in (5.3). While the proposed technique enables us to reconstruct the initial state $x(0)$ from a set of corrupted measurements, it might not always be possible to find such a mapping function. Next, we focus on a larger class of nonlinear systems, and use feedback linearization to transform the nonlinear system in (5.1) into a linear system.

5.2.2 Feedback Linearization

Let us assume that there exist mapping functions $g(y(k))$ and $h_1(x(k))$ (which are not necessarily linear), and a linear map $h_2(e(k))$ such that:

$$f(x(k), e(k)) = g(y(k)) + h_1(x(k)) + h_2(e(k)) \quad (5.4)$$

where $g(y(k)) : \mathbb{R}^p \rightarrow \mathbb{R}^n$, $h_1(x(k)) : \mathbb{R}^n \rightarrow \mathbb{R}^n$, and $h_2(e(k)) : \mathbb{R}^p \rightarrow \mathbb{R}^n$ are non-zero. Without loss of generality, we can choose the control input $u(k)$ such that $u(k) = -h_1(x(k)) + v(k)$. Note that the specific form of our control input does not mean that we cannot use the estimator in control applications. Here, $v(k)$ allows us to choose our control strategy in the desired way (e.g., LQG control). By using this control input, we cancel out the nonlinear term $h_1(x(k))$, and obtain:

$$g(y(k)) = x(k+1) - Ax(k) - v(k) - h_2(e(k)).$$

We can now construct a vector Y as follows:

$$\begin{aligned} Y &= \begin{bmatrix} y(0) \\ y(1) - C(g(y(0)) + v(0)) \\ y(2) - C(Ag(y(0)) + Av(0) + g(y(1)) + v(1)) \\ \vdots \\ y(T-1) - C(A^{T-2}g(y(0)) + A^{T-2}v(0) + \dots) \end{bmatrix} \\ &= \Phi x(0) + \begin{bmatrix} 0 \\ Ch_2(e(0)) \\ CAh_2(e(0)) + Ch_2(e(1)) \\ \vdots \\ CA^{T-2}h_2(e(0)) \dots \end{bmatrix} \end{aligned} \quad (5.5)$$

Note that $h_2(\cdot)$ is a linear map (i.e., $h_2(e(k)) = He(k)$ where $H \in \mathbb{R}^{n \times p}$). Hence, we obtain:

$$Y = \Phi x(0) + \Psi E \quad (5.6)$$

where matrices $\Phi \in \mathbb{R}^{p \times n}$ and $\Psi \in \mathbb{R}^{p \times p \times T}$ are as follows:

$$\Phi = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{T-1} \end{bmatrix}, \Psi = \begin{bmatrix} I & & & \\ CH & I & & \\ CAH & CH & I & \\ \ddots & \ddots & \ddots & \ddots \\ CA^{T-2}H & \dots & \dots & \dots & I \end{bmatrix}.$$

We can now apply the error correction method introduced in Section 4.3 to the linearized system in (5.6) and reconstruct $x(0)$ if the condition in Lemma 2 is satisfied.

In this study, we focus on sensor attack within a noiseless framework. Next, we consider an interconnected power system with several synchronous generators, and illustrate how the proposed nonlinear state estimation approach can be applied for secure state estimation of dynamic states (i.e., generator' phase angles and rotors' speeds).

5.3 Power System State Estimation

We first introduce the physical layer model of an interconnected power system comprising several synchronous generators and buses, and then introduce a graph-theoretic model to describe the communication network which interconnects the wide-area and local controllers of the power system. Figure 5.1 illustrates the interactions between the physical and cyber layers in the system. Note that the components of the system and the notation used in this figure will be introduced throughout this section. Finally, we introduce two categories of cyber attacks that can potentially corrupt measurements and degrade the system's performance.

5.3.1 Physical Layer Model

Consider a power system comprising G generators and B buses. We assume that G of the buses are generator buses, and that the remaining buses ($B - G$ buses) are load buses. Let \mathcal{B} and \mathcal{V} denote the set of buses and transmission lines, respectively. Here, we assume that the corresponding graph $\mathcal{H}(\mathcal{B}, \mathcal{V})$ is connected, and that the network topology is fixed and known.

Load buses: Let V_i and δ_i denote the magnitude and phase angle of the voltage phasor, respectively, at load bus $i \in \mathcal{L}$ where \mathcal{L} is the set of load buses ($|\mathcal{L}| = B - G$). Let P_i^e be the total active power leaving bus i (i.e., the real power drawn by the load at bus i equals $-P_i^e$). P_i^e can be computed by

$$P_i^e = \sum_{j \in \mathcal{B}} V_i V_j |y_{ij}| \sin(\delta_i - \delta_j + \phi_{ij}) \quad (5.7)$$

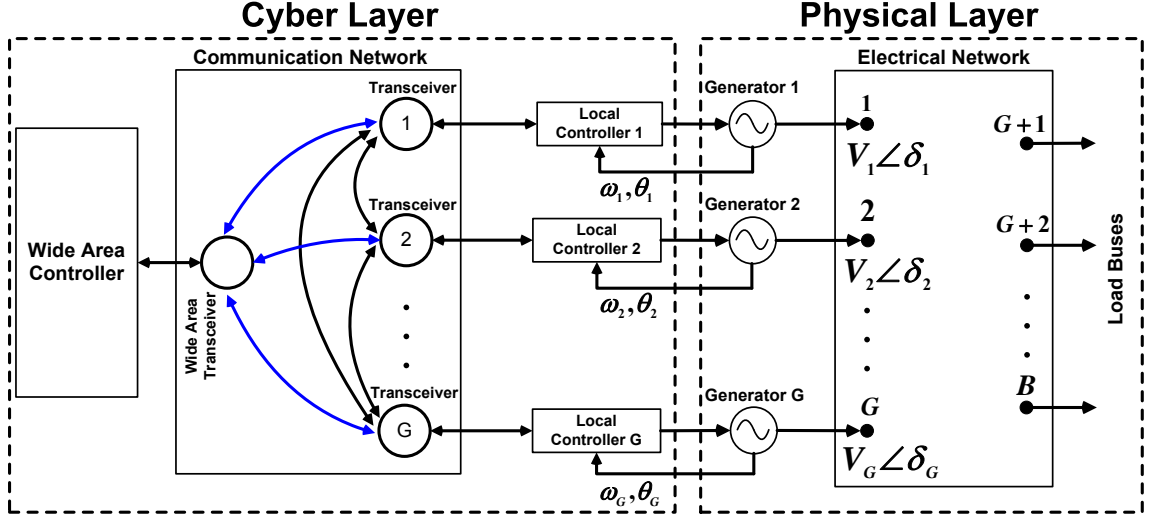


Figure 5.1: A graphical depiction of the power system including both the physical and cyber layers.

where $y_{ij} = g_{ij} + \sqrt{-1}b_{ij}$ is the admittance of the line between buses i and j , and ϕ_{ij} equals $\arctan(g_{ij}/b_{ij})$. Note that $g_{ij} = g_{ji} \geq 0$ and $b_{ij} = b_{ji} > 0$ are the conductance and susceptance of the line between buses i and j , respectively.

Generator buses: Let $\widehat{E}_i = E_i/\theta_i$ denote the internal voltage phasor of the generator connected to bus $i \in \mathcal{G}$ where \mathcal{G} is the set of generator buses in the system. According to the synchronous machine theory, E_i is constant and θ_i is the angular position of the generator rotor as measured with respect to a synchronous reference rotating at the nominal system electrical frequency ω_0 . We assume that the voltages at the generator buses are controlled via droop control, and that all the generator terminal buses are equipped with fast response energy storage units which are controlled via local and wide area controllers. Under these assumptions, for a synchronous generator connected to bus $i \in \mathcal{G}$, the dynamic variables are the generator phase angle θ_i and the rotor electrical angular speed ω_i , and the generator dynamics can be described by [51]

$$\dot{\theta}_i = \omega_i - \omega_0 \quad (5.8)$$

$$\frac{2H_i}{\omega_0} \dot{\omega}_i = P_i^m - P_i^e - \frac{d_i}{\omega_0} (\omega_i - \omega_0) + U_i \quad (5.9)$$

where H_i is the machine inertia constant, d_i is the damping coefficient of the generator, U_i is the external stabilizing energy source at generator bus i , and P_i^m is the mechanical power input to the generator.

Each generator terminal bus $i \in \mathcal{G}$ is equipped with a fast response energy storage, such as flywheels, to improve the system stability. Although synchronous

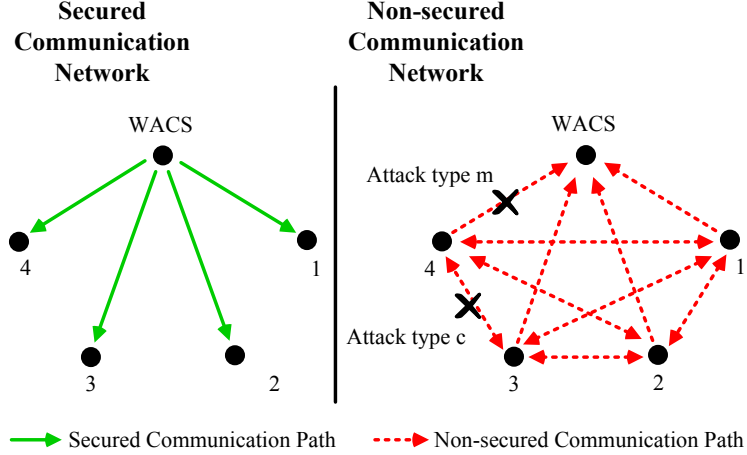


Figure 5.2: A graphical depiction of the cyber network model: For simplicity, we focus on a system with only four generators in this figure. The graph in green (solid lines) shows the secured information flow (i.e., the set of information flows from the WACS to local controllers) while the graph in red (dotted lines) represents the set of non-secured information flows.

generators are typically equipped with local controllers, such as exciter and governor controls, these local controllers only have access to local states and often have slow reaction to rapid system wide perturbations. A local cyber-enabled controller at the generator bus can potentially provide faster response time by using PMU measurements of its neighbors [22], [23].

The energy storage receives a measurement-based control signal computed from PMU measurements, and injects U_i per unit values of power into bus i if $U_i \geq 0$; otherwise, it absorbs U_i per unit values of power from bus i . Similar to the study in [22], we develop a feedback linearization controller, and assume that the local controller at bus $i \in \mathcal{G}$ implements the following feedback linearization control law

$$U_i = -P_i^m + P_{i,\text{meas}}^e - F_i \left(\frac{\omega_i}{\omega_0} - 1 \right) \quad (5.10)$$

where $P_{i,\text{meas}}^e$ is computed locally by the controller at bus i , and $F_i \geq 0$ is a design parameter. For more information on the impact of the parameter F_i on the transient behavior of the system, we refer the reader to [22], [23].

5.3.2 Cyber Layer Model

To maintain the system's stability, the system operator has equipped each generator with a local controller, PMU, and transceiver through which information can be exchanged with the local controllers of other generators as well as the WACS.

These transceivers are connected through a communication network which sends PMU measurements, including rotors' speeds and generators' phase angles, to different transceivers. The communication network, PMUs, and transceivers are not secured, and hence they are subject to cyber attacks and communication failures.

In this study, we assume that the communication paths from the WACS to the local controllers are secured while other communication paths are not secured. Hence, the communication network interconnecting the transceivers can be described by two directed graphs, one for secured information flow and one for non-secured information flow, as shown in Figure 5.2.

Typically, the WACS is strongly protected against cyber attacks, and the local transceivers are more vulnerable to cyber attacks and communication failures than the WACS. For more information, we refer the reader to the North American Electric Reliability Corp. (NERC)s Critical Infrastructure Protection (CIP) standards [69]. In particular, we refer the reader to 1) CIP-002 BES Cyber System Categorization that identifies control centers as a "High Impact Rating", and 2) CIP-005 Electronic security Perimeter(s) and CIP-006 Physical Security of BES Cyber Systems to see what requirements are needed for high impact systems. The CIP Standards explain why we assume that the communication paths from the WACS to the local controllers is secured.

To maintain the system's stability in the presence of attacks and failures, the WACS needs to perform secure state estimation before using the received data (e.g., ω_i 's and θ_i 's) for computing wide area control signals and for monitoring local controllers. To do so, we distinguish two types of attacks:

- **c-attack:** an attack that corrupts communication channels between local controllers.
- **m-attack:** an attack that affects communication channels between a local controller and the WACS.

We assume that at any time instant, the cyber layer is subject to either a c-attack or an m-attack, but not both. This is discussed in detail in Section 5.4.2. However, both of these types of attacks and the set of attacked measurements can change at each time instant. These types of attacks are illustrated in Figure 5.2 for a power system comprising four generators.

Next, by using the proposed secure state estimation technique, we develop a secure state estimator for estimating the dynamic states (i.e., generator' phase angles and rotors' speeds) of the power network.

5.4 Secure State Estimator for Wide Area Control Systems

The system dynamics and power flows can be described by the algebraic differential equations in (5.7)-(5.9). However, in order to use the proposed secure state estimation technique, we need to describe the system by a set of purely differential equations. To do so, we reduce the power network into a network of electro-mechanical oscillators, comprising the G generators, by using the Kron reduction technique². Let \mathcal{V}' denote the set of transmission lines between the G generators after performing the Kron reduction technique, and let $\mathcal{K}(\mathcal{G}, \mathcal{V}')$ denote the corresponding graph. This graph is connected and has $|\mathcal{V}'|$ edges where $|\mathcal{V}'| \leq G(G-1)/2$.

We can now describe the power system by

$$\begin{aligned} \dot{\theta}_i &= \omega_i - \omega_0 \\ \frac{2H_i}{\omega_0} \dot{\omega}_i &= P_i^m - \sum_{j \in \mathcal{N}_i} E_i E_j |\hat{y}_{ij}| \sin(\theta_i - \theta_j + \hat{\phi}_{ij}) \\ &\quad - \frac{d_i}{\omega_0} (\omega_i - \omega_0) + U_i \end{aligned} \quad (5.11)$$

where $\hat{y}_{ij} = \hat{g}_{ij} + \sqrt{-1} \hat{b}_{ij}$ denotes the admittance of the Kron-reduced equivalent line between generators i and j , and $\hat{\phi}_{ij}$ equals $\arctan(\hat{g}_{ij}/\hat{b}_{ij})$. \mathcal{N}_i denotes the set of neighbors of generator i in graph $\mathcal{K}(\mathcal{G}, \mathcal{V}')$ (i.e., the reduced network).

In this study, we assume that the WACS performs a monitoring role and does secure estimation, and consider the mechanical input power P_i^m and the storage control signal U_i as local control signals which are computed based on PMU measurements and wide area control signals (e.g., area control error) [51]. When measurements are attacked, the estimated values of power flows or phase angles will not be equal to the actual values in the system (e.g., $P_{i,\text{meas}}^e \neq P_i^e$), and hence local controllers might send inaccurate signals to physical components. The WACS estimates attacks from measurements and communicates estimated attack signals to each generator. Then, each generator subtracts the estimated attack from the received measurements, as to obtain the most accurate values of ω_i 's and θ_i 's, and to make sure that the local controller will send accurate signals to the physical components that are under its control.

²Kron reduction is a graph-based technique used in power systems to eliminate algebraic load equations and to reduce the order of the interconnections between the synchronous generators [17]. This technique transforms an interconnected power system into an equivalent grid between the synchronous generators of the power system.

5.4.1 Formulation of Secure Estimation

The local controller at generator i computes the control input U_i using (5.10), in which $P_{i,\text{meas}}^e$ is calculated from PMU measurements:

$$P_{i,\text{meas}}^e(t) = \sum_{j \in \mathcal{N}_i} E_i E_j |\hat{y}_{ij}| \sin(y_{ii}^c(t) - y_{ij}^c(t) + \hat{\phi}_{ij}), \quad (5.12)$$

where y_{ij}^c is the measured rotor angle of generator j (i.e., θ_j) received at generator i 's local controller. We assume that these PMU measurements are subject to attack:

$$y_{ij}^c(t) = \theta_j(t) + e_{ij}^c(t), \quad i \in \mathcal{G}, \quad j \in \mathcal{N}_i \cup \{i\}, \quad (5.13)$$

where e_{ij}^c represents the attack signal. As mentioned earlier, we refer to this as a c-attack (see Figure 5.2). In addition, we assume $e_{ii}^c(t) = 0$ for all t . Note that θ_i is measured locally, and therefore it is not subject to cyber attack, i.e., $y_{ii}^c(t) = \theta_i(t)$ for all t .

To perform secure estimation, the WACS receives measurements y_{ij}^m from all the local controllers. Since we assume the communication flows from the local controllers to the WACS are not secured, measurements y_{ij}^m can be subject to attack:

$$y_{ij}^m(t) = y_{ij}^c(t) + e_{ij}^m(t), \quad i \in \mathcal{G}, \quad j \in \mathcal{N}_i \cup \{i\} \quad (5.14)$$

where e_{ij}^m represents the corruption in y_{ij}^c . We refer to these attacks as m-attacks (see Figure 5.2).

We now apply the forward Euler discretization scheme to this continuous-time system and obtain the following discrete-time approximation, assuming a constant discretization step T_s for all k :

$$\begin{aligned} \theta_i(k+1) &= \theta_i(k) + T_s(w_i(k) - \omega_0) \\ \omega_i(k+1) &= \alpha \omega_i(k) + \beta \\ &\quad + \sum_{j \in \mathcal{N}_i} f_{ij}(\theta_i(k), \theta_j(k), y_{ii}^c(k), y_{ij}^c(k)) \end{aligned} \quad (5.15)$$

where $\alpha = 1 - \frac{T_s(d_i + F_i)}{2H_i}$, $\beta = \frac{T_s \omega_0(d_i + F_i)}{2H_i}$, $f_{ij}(\cdot) = \tilde{G}_{ij}[\sin(\theta_i(k) - \theta_j(k) + \hat{\phi}_{ij}) - \sin(y_{ii}^c(k) - y_{ij}^c(k) + \hat{\phi}_{ij})]$ and $\tilde{G}_{ij} = -\frac{T_s \omega_0 E_i E_j |\hat{y}_{ij}|}{2H_i}$.

Using (5.13) and (5.14), $f_{ij}(\cdot)$ can be re-written in terms of y_{ij}^m 's, which are the measurements received at the WACS, as follows:

$$\begin{aligned} f_{ij}(\cdot) &= \tilde{G}_{ij}[\sin(\hat{\phi}_{ij} + \theta_i(k) - \theta_j(k)) \\ &\quad - \sin(\hat{\phi}_{ij} + y_{ii}^c(k) - y_{ij}^c(k))] \\ &= \tilde{G}_{ij} \sin(\hat{\phi}_{ij} + y_{ii}^m(k) - y_{ij}^m(k) - e_{ii}^m(k) \\ &\quad + e_{ij}^m(k) + e_{ij}^c(k)) - \tilde{G}_{ij} \sin(\hat{\phi}_{ij} + y_{ii}^m(k) \\ &\quad - y_{ij}^m(k) - e_{ii}^m(k) + e_{ij}^m(k)) \\ &= G_{ij}^s(k) \epsilon_{ij}^c(k) - G_{ij}^c(k) \epsilon_{ij}^s(k), \end{aligned}$$

where $G_{ij}^s(k) = \tilde{G}_{ij} \sin(\hat{\phi}_{ij} + y_{ii}^m(k) - y_{ij}^m(k))$, $G_{ij}^c(k) = \tilde{G}_{ij} \cos(\hat{\phi}_{ij} + y_{ii}^m(k) - y_{ij}^m(k))$ are known to the WACS. On the other hand, ϵ_{ij}^c and ϵ_{ij}^s are functions of unknown attack signals and are defined as:

$$\begin{aligned}\epsilon_{ij}^c(k) &= \cos(e_{ii}^m(k) - e_{ij}^m(k) - e_{ij}^c(k)) \\ &\quad - \cos(e_{ii}^m(k) - e_{ij}^m(k)) \\ \epsilon_{ij}^s(k) &= \sin(e_{ii}^m(k) - e_{ij}^m(k) - e_{ij}^c(k)) \\ &\quad - \sin(e_{ii}^m(k) - e_{ij}^m(k)).\end{aligned}\tag{5.16}$$

In other words, $f_{ij}(\cdot)$ is now a linear function of the unknowns: $\epsilon_{ij}^c(k)$ and $\epsilon_{ij}^s(k)$, whose coefficients can be computed by the WACS from the received measurements. In addition, if there is no attack on any of the communication channels in the system at time slot k , then $\epsilon_{ij}^c(k) = \epsilon_{ij}^s(k) = 0$.

The state space model of the i -th generator is given by:

$$\begin{aligned}x_i(k+1) &= A_i x_i(k) + q_i + H_i(k) \epsilon_i(k) \\ &= \begin{bmatrix} 1 & T_s \\ 0 & \alpha \end{bmatrix} x_i(k) + \begin{bmatrix} -T_s \omega_0 \\ \beta \end{bmatrix} + \begin{bmatrix} 0 \\ h_i(k)^\top \end{bmatrix} \epsilon_i(k)\end{aligned}\tag{5.17}$$

where the state vector $x_i(k) = [\theta_i(k), \omega_i(k)]^\top$ and

$$\begin{aligned}h_i(k)^\top &= [G_{i\mathcal{N}_i(1)}^s(k), \dots, G_{i\mathcal{N}_i(l_i)}^s(k), \\ &\quad G_{i\mathcal{N}_i(1)}^c(k), \dots, G_{i\mathcal{N}_i(l_i)}^c(k)] \in \mathbb{R}^{1 \times 2l_i} \\ \epsilon_i(k) &= [\epsilon_{i\mathcal{N}_i(1)}^c(k), \dots, \epsilon_{i\mathcal{N}_i(l_i)}^c(k), \\ &\quad \epsilon_{i\mathcal{N}_i(1)}^s(k), \dots, \epsilon_{i\mathcal{N}_i(l_i)}^s(k)]^\top \in \mathbb{R}^{2l_i \times 1}.\end{aligned}$$

Here, $\mathcal{N}_i(j)$ is the j -th generator in the neighborhood of generator i and l_i is the cardinality of the set \mathcal{N}_i .

Consider the enlarged system with G generators in the network:

$$\begin{aligned}X(k+1) &= AX(k) + q + H(k)\epsilon(k) \\ Y(k) &= CX(k) + DE(k)\end{aligned}\tag{5.18}$$

where

$$\begin{aligned}
X(k) &= [x_1(k); \dots; x_G(k)] \in \mathbb{R}^{2G \times 1} \\
A &= \text{blkdiag}\{A_1, \dots, A_G\} \in \mathbb{R}^{2G \times 2G} \\
q &= [q_1; \dots; q_G] \in \mathbb{R}^{2G \times 1} \\
H(k) &= \text{blkdiag}\{H_1(k), \dots, H_G(k)\} \in \mathbb{R}^{2G \times 4L} \\
\epsilon(k) &\triangleq [\epsilon_1(k); \dots; \epsilon_G(k)] \in \mathbb{R}^{4L \times 1} \\
Y(k) &= [Y_i(k); \dots; Y_G(k)] \in \mathbb{R}^{(G+2L) \times 1} \\
Y_i(k) &= [y_{ii}(k); y_{i\mathcal{N}_i(1)}(k); \dots; y_{i\mathcal{N}_i(l_i)}(k)] \in \mathbb{R}^{(1+l_i) \times 1} \\
D &= [D_1, D_2] \in \mathbb{R}^{(G+2L) \times (G+4L)} \\
D_1 &= \text{blkdiag} \left\{ \begin{bmatrix} 0 \\ I_{l_1, l_1} \end{bmatrix}, \dots, \begin{bmatrix} 0 \\ I_{l_G, l_G} \end{bmatrix} \right\} \in \mathbb{R}^{(G+2L) \times 2L} \\
D_2 &= I_{G+2L, G+2L} \in \mathbb{R}^{(G+2L) \times (G+2L)} \\
E(k) &= [E_1^c(k); \dots; E_G^c(k); E_1^m(k); \dots; E_G^m(k)] \\
&\in \mathbb{R}^{(G+4L) \times 1} \\
E_i^c(k) &= [e_{i\mathcal{N}_i(1)}^c(k); \dots; e_{i\mathcal{N}_i(l_i)}^c(k)] \in \mathbb{R}^{l_i \times 1} \\
E_i^m(k) &= [e_{ii}^m(k); e_{i\mathcal{N}_i(1)}^m(k); \dots; e_{i\mathcal{N}_i(l_i)}^m(k)] \in \mathbb{R}^{(1+l_i) \times 1}
\end{aligned}$$

and $L = \frac{\sum_i l_i}{2}$ represents the total number of edges / links in the network. Matrix $C \in \mathbb{R}^{(G+2L) \times 2G}$ is given as follows: let the a -th element of vector Y be y_{ij}^m , then the (a, b) -th entry of C is given by

$$C_{(a,b)} = \begin{cases} 1 & \text{if } 2j - 1 = b \\ 0 & \text{otherwise.} \end{cases}$$

Consider T time steps of measurements (i.e., $k = \{0, \dots, T-1\}$) and define:

$$\bar{Y} = \begin{bmatrix} Y(0) \\ Y(1) - Cq \\ \vdots \\ Y(T-1) - C \sum_{m=0}^{T-2} A^{T-2-m} q \end{bmatrix} \in \mathbb{R}^{(G+2L)T \times 1} \quad (5.19)$$

then

$$\bar{Y} = \Phi X(0) + \Psi \bar{E} \quad (5.20)$$

where $\Phi = [C; CA; \dots; CA^{T-1}] \in \mathbb{R}^{(G+2L)T \times 2G}$ is the T -step observability matrix of the system, $\bar{E} = [E(0); \dots; E(T-1); \epsilon(0); \dots; \epsilon(T-2)] \in \mathbb{R}^{((G+4L)T + 4L(T-1)) \times 1}$ and

$\Psi = [\Psi_1 \ \Psi_2]$, with $\Psi_1 \in \mathbb{R}^{(G+2L)T \times (G+4L)T}$ and $\Psi_2 \in \mathbb{R}^{(G+2L)T \times 4L(T-1)}$ as follows:

$$\Psi_1 = \text{blkdiag}\{D, \dots, D\}$$

$$\Psi_2 = \begin{bmatrix} 0 & 0 & \dots \\ CH(0) & 0 & \dots \\ CAH(0) & CH(1) & \dots \\ \vdots & \vdots & \ddots \\ CA^{T-2}H(0) & \dots & CH(T-2) \end{bmatrix}.$$

We can choose $\Omega \in \mathbb{R}^{((G+2L)T-2G) \times (G+2L)T}$ such that $\Omega\Phi = 0$, then:

$$\tilde{Y} = \Omega\bar{Y} = \Omega\Psi\bar{E}, \quad (5.21)$$

where $\Omega\Psi \in \mathbb{R}^{((G+2L)T-2G) \times ((G+4L)T+4L(T-1))}$.

5.4.2 Challenges in Secure Estimation due to the Power System's Dynamics

The linear system in (5.21) is in the form of (4.8). Hence, from Lemma 1, \bar{E} has a unique s -sparse solution if all subsets of $2s$ columns of $\Omega\Psi$ are linearly independent. We now explain that this is not the case in the power systems example as some columns of Ψ are linearly dependent. Let us begin with the following: consider a matrix-vector multiplication $M \cdot v$, where $M = [m_1, \dots, m_n] \in \mathbb{R}^{l \times n}$ and m_i is the i -th column of M , $v = [v_1, \dots, v_n]^T \in \mathbb{R}^{n \times 1}$, and v_i is the i -th entry of v . In the sequel, the phrase “the column of M that corresponds to v_i ” refers to the column of M that multiplies the v_i entry in the matrix-vector multiplication, i.e., m_i .

We now explain why Ψ_2 is rank deficient. Observe that for all i and k , the first row of $H_i(k)$ is equal to zero. Therefore given any matrix $M = [m_1 \ m_2] \in \mathbb{R}^{l \times 2}$, where m_1 and m_2 are the columns of M , we have:

$$\begin{aligned} \text{rank}(M \cdot H_i(k)) &= \text{rank}\left([m_1 \ m_2] \cdot \begin{bmatrix} 0 & 0 & \dots \\ h_{21} & h_{22} & \dots \end{bmatrix}\right) \\ &= \text{rank}([h_{21}m_2 \ h_{22}m_2 \ \dots]) = 1. \end{aligned} \quad (5.22)$$

Since $H(k)$ is block diagonal, we can show that Ψ_2 is also rank deficient. Next, from (5.13) and (5.14), we have $y_{ij}^m(k) = \theta_j(k) + e_{ij}^c(k) + e_{ij}^m(k)$, which means for a given (i, j) -pair ($i \neq j$) and a given time slot k , the two columns of Ψ_1 that correspond to the two terms $e_{ij}^c(k)$ and $e_{ij}^m(k)$ in \bar{E} are identical, i.e., linearly dependent. Therefore, by Lemma 1, the solution \bar{E} obtained by solving (5.21) (i.e., the estimation algorithm introduced in Section 4.3) is not unique. Does this mean we can not uniquely recover the attack signal? A closer analysis reveals the specific entries in \bar{E} that cannot be uniquely identified, and shines light on how to overcome this challenge. Our observations are as follows:

1. Observe from (5.21) that Ψ_2 multiplies the $\epsilon(k)$ terms in \bar{E} , i.e., $(\epsilon(0), \dots, \epsilon(T-2))$. Linear dependence of the columns of Ψ_2 causes the $\epsilon(k)$ terms to be unidentifiable. However, $\epsilon(k)$ can be computed from $E(k)$ using (5.16). In other words, although the $\epsilon(k)$ terms in the solution \bar{E} are not unique, as long as the $E(k)$ terms in \bar{E} are unique, then we can determine the $\epsilon(k)$ terms uniquely using (5.16).
2. The identical columns of Ψ_1 that correspond to the two terms $e_{ij}^c(k)$ and $e_{ij}^m(k)$ in \bar{E} means that it is only possible to uniquely identify the sum $e_{ij}^c(k) + e_{ij}^m(k)$, but not the individual terms: $e_{ij}^c(k)$ and $e_{ij}^m(k)$. To overcome this challenge, we make the assumption that at any time k , the system is subject to either a c-attack or an m-attack, but not both. In other words, $e_{ij}^m(k)$ and $e_{ij}^c(k)$ cannot both be non-zero, thus making them identifiable.

Next, we explain our secure estimation algorithm.

5.4.3 Assumptions and Secure Estimation with 2-Step Delay

As mentioned earlier, we assume that at any time slot k , the cyber layer is subject to either a c-attack or an m-attack, but not both at the same time. However, both the types of attacks and the set of attacked measurements can change at each time step. In addition, the WACS does not know *a priori* which type of attack the network is subjected to. Hence, secure estimation techniques are required to determine the type of attack, as well as the exact corruption signals.

Using the difference equation in (5.18), we find that

$$\begin{aligned} f_{2\text{-step}}(\epsilon(k-2), E(k)) &= Y(k) - CA^2 \cdot X(k-2) - CAq \\ &\quad - Cq - CA \cdot H(k-2) \cdot \epsilon(k-2) - DE(k) = 0 \end{aligned} \quad (5.23)$$

where the first equality uses $CH(k) = 0$ for all k . Observe that if it is an m-attack at time k , then $e_{ij}^c(k) = 0$ and $e_{ij}^c(k) + e_{ij}^m(k) = e_{ij}^m(k)$, furthermore, $\epsilon(k) = 0$. On the other hand, if it is a c-attack at time k , then $e_{ij}^m(k) = 0$ and $e_{ij}^c(k) + e_{ij}^m(k) = e_{ij}^c(k)$, in addition, $\epsilon(k) \neq 0$. Combining these two observations with (5.23), we propose the following algorithm which can be used by the WACS to determine the type of attack and the exact corruption signals, with a 2-step delay.

We first introduce some notation used in the algorithm. Let $E_b(k)$ denote the estimated vector $E(k)$ without imposing the assumption that only a c-attack or an m-attack can occur. $E_{c\text{-att}}(k)$ and $E_{m\text{-att}}(k)$ denote the estimated vector $E(k)$ if it is a c-attack or an m-attack at time k , respectively, and can be computed from $E_b(k)$. For example, to obtain $E_{c\text{-att}}(k)$, we set all e_{ij}^m terms in $E_{c\text{-att}}(k)$ to zero, and set all e_{ij}^c terms equal to the sum of corresponding e_{ij}^c and e_{ij}^m terms in $E_b(k)$. We can obtain $E_{m\text{-att}}(k)$ in a similar fashion. Finally, $\epsilon_{c\text{-att}}(k)$ and $\epsilon_{m\text{-att}}(k)$ are the $\epsilon(k)$ vectors

Algorithm 3 Secure Estimation

- 1: **for** each k **do**
 - 2: Estimate $\bar{E}(k)$ by solving the following l_1 -minimization problem:

$$\bar{E}_b(k) = \arg \min \|\bar{E}\|_{l_1} \text{ subject to } \tilde{Y} = \Omega\Psi\bar{E}.$$
 - 3: Extract $E_b(k-2)$ and $E_b(k)$ from $\bar{E}_b(k)$.
 - 4: Evaluate equation (5.23) for the case with an m-attack at time $k-2$, using the observation that $\epsilon_{m\text{-att}}(k-2) = 0$.
 - 5: Evaluate equation (5.23) for the case with a c-attack at time $k-2$, by first computing $E_{c\text{-att}}(k-2)$ and then use (5.16) to obtain $\epsilon_{c\text{-att}}(k-2)$.
 - 6: **if** $\|f_{2\text{-step}}(\epsilon_{c\text{-att}}(k-2), E_b(k))\| < \|f_{2\text{-step}}(\epsilon_{m\text{-att}}(k-2), E_b(k))\|$ **then**
 - 7: It is a c-attack at $k-2$: $E(k-2) = E_{c\text{-att}}(k-2)$ and $\epsilon(k-2) = \epsilon_{c\text{-att}}(k-2)$
 - 8: **else**
 - 9: It is an m-attack at $k-2$: $E(k-2) = E_{m\text{-att}}(k-2)$ and $\epsilon(k-2) = \epsilon_{m\text{-att}}(k-2) = 0$
 - 10: **end if**
 - 11: **end for**
-

computed from $E_{c\text{-att}}(k)$ and $E_{m\text{-att}}(k)$, respectively. With these notations in hand, we now present our estimation algorithm in Algorithm 3.

To summarize, as a result of the system dynamics and the proposed model, it is not possible to recover the exact corruption if the system is subjected to both c- and m-attacks at the same time. In light of this, we make the simplifying assumption that at any time k , the system may only be subject to one type of attack. However, the type of attack can change over time. Then, by comparing the actual measurements with the system trajectories that would result from each type of attack, we can determine both the attack type and the exact corruption signals, with a 2-step delay. Note that at time k , this secure state estimation algorithm is able to detect the presence of attacks at times $k-1$ and k , merely not the exact attack signals. Next, we numerically demonstrate the effectiveness of the proposed state estimation algorithm.

5.5 Numerical Example

We focus on the New England power system comprising 10 generators and 39 buses, and simulate the system for $t = 20$ seconds with a discretization step of $1/60$ seconds. The values of the system parameters are taken from [3], [74]. The power system is running under normal condition from $t = 0$ to $t = 2$ seconds. At $t = 2$ seconds, a three-phase fault occurs at Bus 17. Then, Line 17-18 is tripped out to clear the fault. However, the WACS is unaware of this fault at this time. Two seconds later, at $t = 4$ seconds, the WACS detects the occurrence of this fault, i.e., there is a

2 second-delay in the WACS being able to respond to the fault. We conduct load flow analysis of the power system before and after the occurrence of the 3-phase fault, to find the values of P_i^e , θ_i , and $|E_i|$ for each generator. We demonstrate the effectiveness of our proposed secure estimation method through simulations of 3 different scenarios:

1. *Scenario 1:* There is no simultaneous cyber attack on the power system.
2. *Scenario 2:* The power system is also under cyber attack, and it is not protected by secure estimation.
3. *Scenario 3:* The power system is also under cyber attack, and it is protected by secure estimation.

The plots titled “No Attack” in Figure 5.3 show the simulation results of Scenario 1: an attack-free power system under a three-phase fault. For clarity, only phase angles and rotor speeds of generators 1, 2 and 3 are shown in Figure 5.3. At $t = 0$ seconds, the system is under equilibrium, all ten generators’ rotor speeds are at the nominal value, ω_0 , of 60 Hz, and their phase angles are 6.85° , 5.09° , 6.28° , 8.81° , 7.38° , 11.30° , 14.74° , 8.35° , 7.63° and -13.11° , respectively. At $t = 2$ seconds, a three-phase fault occurs at Bus 17 which causes a change in the line admittances (y_{ij} ’s and ϕ_{ij} ’s) and consequently, the total active power leaving bus i , P_i^e . However, the WACS is unaware of this fault until $t = 4$ seconds. During this 2 second-delay, the WACS is unaware of the fault and continues to use the pre-fault line admittance values in the secure estimation algorithm. The local controllers at the generators continue to compute the control input U_i using the received measurements from the WACS, which leads to a mismatch between $P_{i,\text{meas}}^e$ and P_i^e , and causes the phase angles and rotor speeds of the generators to deviate from their equilibrium. At $t = 4$ seconds, the WACS becomes aware of the fault. It computes the new line admittance values under this fault, and uses the new line admittance values in the secure estimation algorithm. The local controllers then use the received estimates to compute the local control input U_i , making $P_{i,\text{meas}}^e = P_i^e$ again. As a result, the generators’ rotor speeds slowly converge back to 60 Hz and their rotor angles settle at new equilibrium values.

In Scenarios 2 and 3, in addition to the 3-phase fault, the power system is also subject to the following cyber attack. Malicious attacks targeted at generator 1 are injected from $t = 0.33$ seconds onwards. A set of 10 measurements that varies with time are corrupted. More specifically, at each time step, the attacker randomly chooses to perform either a c-attack or an m-attack. In the case of a c-attack, the attacker corrupts phase angle measurements that generator 1 receives from all its 9 neighbors (i.e., $y_{1,2}^c, y_{1,3}^c, \dots, y_{1,10}^c$) with independent Gaussian signals from the distribution $\mathcal{N}(0, 180^\circ)$. In addition, a constant signal of 90° is injected into the measurement $y_{1,2}^c$. In the case of an m-attack, the attacker randomly chooses 9 measurements from the set of 10 measurements that generator 1 submits to the WACS (i.e., $y_{1,1}^m, y_{1,2}^m, \dots, y_{1,10}^m$), and corrupts each chosen measurement with an independent

Gaussian signal from $\mathcal{N}(0, 180^\circ)$. Similarly, an additional constant signal of 90° is injected into the measurement $y_{1,2}^m$. The left plot in Figure 5.4 shows the true attack signals. Rows 1 to 9 correspond to c-attacks: $e_{1,2}^c, e_{1,3}^c, \dots, e_{1,10}^c$, and rows 10 to 19 correspond to m-attacks: $e_{1,1}^m, e_{1,2}^m, \dots, e_{1,10}^m$. Since constant signals of 90° are injected on top of the Gaussian attacks to $y_{1,2}^c$ or $y_{1,2}^m$ for c-attacks or m-attacks, respectively, the mean attack signals are higher for row 1 (i.e., $e_{1,2}^c$) and row 11 (i.e., $e_{1,2}^m$). For clarity, the measurements that are not attacked during the simulation are not shown.

In Scenario 2, no secure estimation-based protection is implemented. Therefore, when the system is under cyber attack, the local controller at generator 1 computes $P_{1,\text{meas}}^e$ using corrupted measurements, causing $P_{1,\text{meas}}^e \neq P_1^e$. As a result, the feedback control law in (5.10) fails to linearize the system dynamics (5.9). The constant signal of 90° injected on top of the Gaussian attacks causes oscillations in the rotor speed of generator 1 due to the sine term in its dynamics (refer to Equation (5.10) to (5.13)). The oscillations in the rotor speed then leads to oscillations in generator 1's rotor angle. The plots titled “Under Attack, no SE” in Figure 5.3 show that these oscillations are observed on top of the system's response to the 3-phase fault, and prevents generator 1's phase angle to reach a new equilibrium even after the fault has cleared. In addition, the cyber attack causes larger differences in other generators' equilibrium rotor angles before and after the fault. For example, in Scenario 1, when there is no cyber attack, generator 2 and 3's rotor angles after the fault are -14° and -8° respectively. On the other hand, in Scenario 2, their post-fault equilibrium rotor angles are -25° and -17° respectively.

Finally, in Scenario 3, the power system is subject to the same cyber attack as in Scenario 2. However, the WACS uses secure estimation to protect the system against such attacks. The center and right plots in Figure 5.4 show the secure estimator's estimated attack signal and the estimation error respectively. The results show that the secure estimator correctly estimates the attack signal before the fault happens at $t = 2$ seconds. Between $t = 2$ and $t = 4$ seconds, there are small estimation errors due to model mismatch as the WACS is unaware of the fault and continues to use the pre-fault line admittance values in the secure estimation algorithm. However, once the WACS is informed of the fault at $t = 4$ seconds, the model mismatch is removed and estimation error is cleared. The estimated the attack signals are then subtracted from the corrupted measurements to recover the true rotor angles and speeds. The reconstructed measurements are communicated to all generators, and used to compute $P_{1,\text{meas}}^e$. By doing this, the local controllers obtain a value of $P_{1,\text{meas}}^e$ that is a more accurate estimate of the true P_1^e than when no secure estimation was used. The bottom plots, titled “Under Attack, with SE”, in Figure 5.3 show the rotor angles and speeds of generators 1, 2, and 3 in this scenario. Observe that throughout the simulation, generator 1's rotor angle is much more stable in this scenario than in Scenario 2. In addition, note that when the power system is under cyber attack, the behavior of the system with secure estimation (Scenario 3) resembles more closely the system's behavior when there is no cyber attack (Scenario 1), than the system

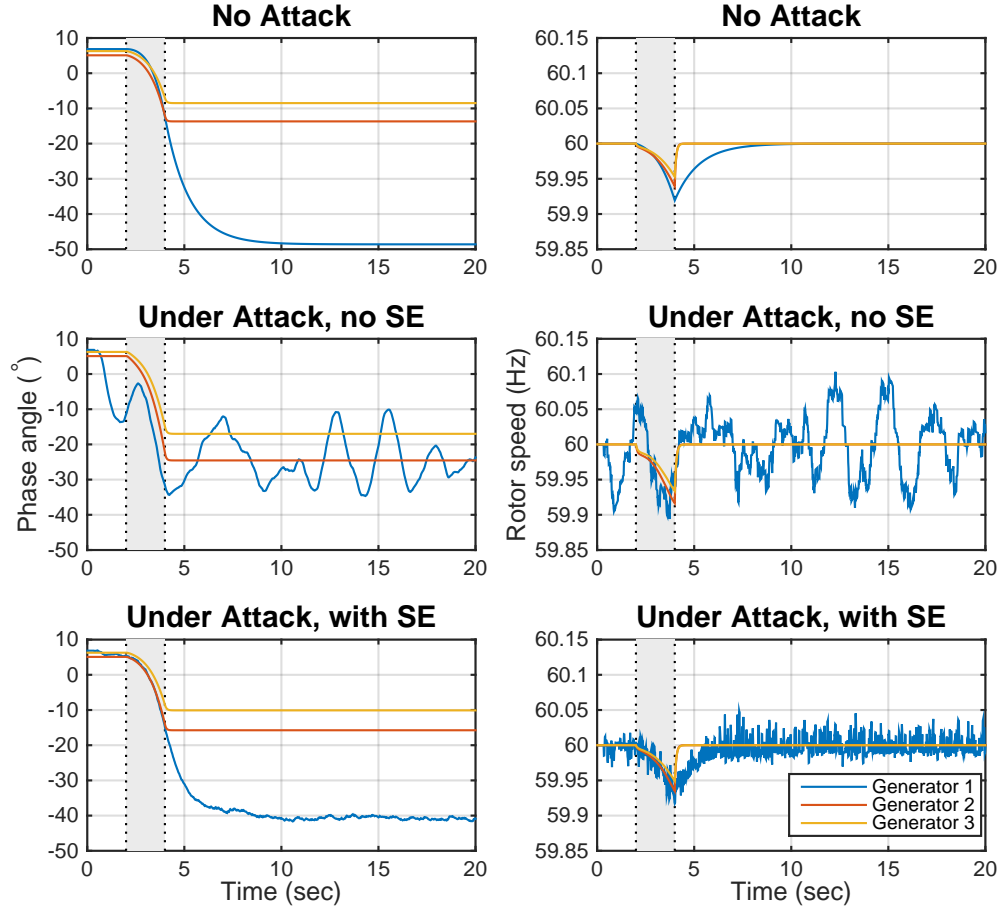


Figure 5.3: Evolution of phase angles and rotor speeds of generators 1, 2 and 3 under a 3-phase fault, in three scenarios: (1) there is no attack, (2) system is under attack and there is no secure estimation (SE), (3) system is under attack and WACS uses SE. Fault happens at $t = 2$ seconds. Grey region marks the 2 seconds delay in WACS being informed of the fault. In (2), cyber attack causes generator 1's rotor angle and speed to oscillate. In (3), incorporating SE damps the large oscillations and makes the system's response more closely resemble that of (1).

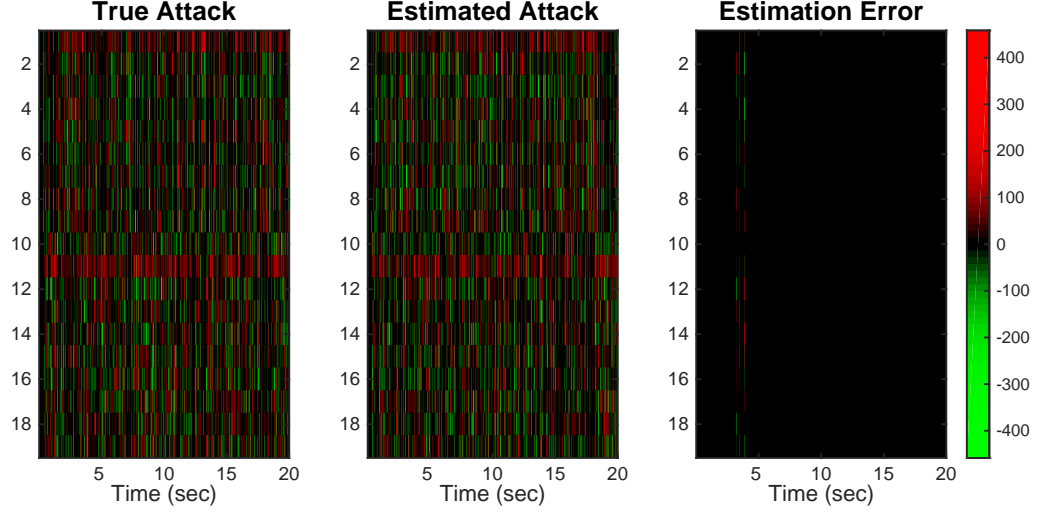


Figure 5.4: True and estimated attack signals: The rows and columns correspond to attacked measurements and time steps, respectively. In subfigures “True Attack” and “Estimated Attack”, the color indicates the attack signal: red is a positive attack, green a negative attack, and black is no attack. In subfigure “Estimation Error”, the black color indicates there is zero estimation error for all measurements at all times.

without secure estimation (Scenario 2) does.

As mentioned earlier, there is a 2-step estimation delay in the proposed secure state estimator. Our numerical results show that this estimation delay will not affect the phase angle and rotor speeds significantly. This can be explained by the fact that the attack signals effect on the system dynamics are scaled by the matrix H (see Equations (26) and (27)) whose entries are very small due to the small discretization time step (1/60 seconds) and the large generators’ inertia (i.e., attack signals cannot immediately have a significant effect on the generators’ phase angles and speeds). The simulation was repeated using a larger discretization time step of 1/30 seconds and the same observations were made (the control feedback gain F_i was adjusted accordingly). Due to space limitations, we do not show the results here.

5.6 Conclusion

We propose a secure state estimator for two classes of nonlinear dynamical systems. We then focus on the wide area control of power systems, and develop an estimator for dynamic states in power systems under cyber-physical attacks and communication failures. Finally, we numerically show that the performance of the cyber layer in power systems can be significantly improved by using our estimator.

Chapter 6

Conclusions and Future Work

There has been a tremendous amount of progress in the development of smart and reliable power systems over the past decade. However, the increased penetration of renewable energy sources and advanced instruments such as PMUs, as part of this movement, also introduced new challenges. This thesis describes initial progress in the path towards a more reliable power system, in particular, by exploiting demand side flexibility by using commercial buildings to provide regulation for the grid frequency, and by using secure state estimation to protect the power system against cyber attacks.

There are many exciting areas of future research in this field. A few high level directions are given below.

Exploring alternative flexibility in buildings: There are a number of alternative flexible loads such as chillers and heat pumps that present great potential for frequency regulation. The flexibility from different loads can be combined to provide a larger regulation capacity, as well as to offer regulation at a wider frequency range.

Frequency regulation from an aggregation of buildings: Power system operators such as PJM and California ISO require a resource to provide a minimum of 0.1 MW of regulation capacity in order to participate in the frequency regulation market. It is unlikely that a single building can satisfy this requirement. One solution is to aggregate several buildings and offer their combined capacity to the regulation market. There has been some initial theoretical work in how to design the contract in this scenario [5]. With the experimental setup presented in this thesis, the feasibility of this idea can now be verified experimentally.

Applying secure estimation methods to other systems: In this thesis, the development of the secure state estimation methods assumes general system dynamics. Therefore the resulting estimation algorithms are applicable to a wide variety of systems, ranging from autonomous ground and aerial vehicles to large systems such

as traffic and water networks.

Reducing computational complexity of the secure estimator: The computational complexity increases with the time index of the estimator. Alternatives such as computing an exact solution to the l_1 -minimization problem in a recursive way may significantly reduce the time required to obtain a new estimate.

Secure estimation for general nonlinear systems: The secure estimation method presented in Chapter 5 focuses on two classes of nonlinear dynamical system. It is an initial attempt at tackling the nonlinear system's secure estimation problem.



Hardware implementation: The secure state estimation methods that we develop can be, and should be validated on hardware platforms.

Bibliography

- [1] Amazon. Amazon Prime Air. <http://www.amazon.com/b?node=8037720011>.
- [2] Anil Aswani, Neal Master, Jay Taneja, Virginia Smith, Andrew Krioukov, David Culler, and Claire Tomlin. Identifying Models of HVAC Systems Using Semiparametric Regression. *American Control Conference*, 2012.
- [3] T. Athay, R. Podmore, and S. Virmani. A practical method for the direct analysis of transient stability. *IEEE Transactions on Power Apparatus and Systems*, 98:573–584, 1979.
- [4] F. Baccino, F. Conte, S. Massucco, F. Silvestro, and S. Grillo. Frequency Regulation by Management of Building Cooling Systems through Model Predictive Control. *Power Systems Computation Conference*, pages 1–7, 2014.
- [5] Maximilian Balandat, Frauke Oldewurtel, Mo Chen, and Claire Tomlin. Contract Design for Frequency Regulation by Aggregations of Commercial Buildings. *52nd Annual Allerton Conference on Communication, Control, and Computing*, 2014.
- [6] H.M. Beides and G.T. Heydt. Dynamic state estimation of power system harmonics using kalman filter methodology. *IEEE Transactions on Power Delivery*, 6(4):1663–1670, 1991.
- [7] Patrick Bouffard. On-board model predictive control of a quadrotor helicopter: Design, implementation, and experiments. Technical Report UCB/EECS-2012-241, University of California Berkeley, December 2012.
- [8] James Braun, Donghun Kim, Miroslav Baric, Pengfei Li, Satish Narayanan, Shui Yuan, Eugene Cliff, John Burns, and Bill Henshaw. Whole building control system design and evaluation: simulation-based assessment. Technical report, GPIC Energy Efficient Buildings Hub, 2012.
- [9] Emmanuel Candes, Justin Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

- [10] Emmanuel Candes and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.
- [11] Alvaro A Cardenas, Saurabh Amin, and Shankar Sastry. Secure control: towards survivable cyber-physical systems. *28th international conference on distributed computing systems workshop*, pages 495–500, June 2008.
- [12] Aranya Chakraborty and Pramod Khargonekar. Introduction to wide-area control of power systems. *Proceedings of American Control Conference*, 2013.
- [13] Wesley J. Cole, Elaine T. Hale, and Thomas F. Edgar. Building Energy Model Reduction for Model Predictive Control Using OpenStudio. *American Control Conference*, 2013.
- [14] Stephen Dawson-Haggerty, Xiaofan Jiang, Gilman Tolle, Jorge Ortiz, and David Culler. SMAP - A Simple Measurement and Actuation Profile for Physical Information. *Proceedings of the 8th International Conference on Embedded Networked Sensor Systems (SenSys)*, 2010.
- [15] Stephen Dawson-Haggerty, Andrew Krioukov, and David E. Culler. Experiences Integrating Building Data with sMAP. Technical report, University of California, Berkeley, 2012.
- [16] Justin R. Dobbs. A Comparison of Thermal Zone Aggregation Methods. *51st Conference on Decision and Control*, 2012.
- [17] Florian Dorfler and Francesco Bullo. Kron reduction of graphs with applications to electrical networks. *IEEE Transactions on Circuits and Systems I: Regular Papers*, pages 150–163, 2012.
- [18] M. Duffy, M. Hiller, D. Bradley, W. Keilholz, and J. Thornton. TRNSYS - Features and Functionality for Building Simulation. *IBSPA Conference*, pages 1950 – 1954, 2009.
- [19] FAA. Press release – DOT and FAA propose new rules for small unmanned aircraft systems. http://www.faa.gov/news/press_releases/news_story.cfm/?newsId=18295. Accessed: 2015-02-15.
- [20] Luca Fabietti, Tomasz T. Gorecki, Faran A. Qureshi, Altug Bitlislioglu, Ioannis Lymperopoulos, and Colin N. Jones. Experimental Implementation of Frequency Regulation Services Using Commercial Buildings. *IEEE Transactions on Smart Grid*, 2016.
- [21] Shaun M. Fallat and Charles R. Johnson. *Totally nonnegative matrices*. Princeton University Press, 2011.

- [22] Abdallah Farraj, Eman Hammad, and Deepa Kundur. A cyber-enabled stabilizing control scheme for resilient smart grid systems. *IEEE Transactions on Smart Grid*, 7(4), 2015.
- [23] Abdallah Farraj, Eman Hammad, and Deepa Kundur. On the use of energy storage systems and linear feedback optimal control for transient stability. *IEEE Transactions on Industrial Informatics*, 13(4), 2017.
- [24] Hamza Fawzi, Paulo Tabuada, and Suhas Diggavi. Secure estimation and control for cyber-physical systems under adversarial attacks. *IEEE Transactions on Automatic Control*, 59(6):1454–1467, 2014.
- [25] Miao Fei, Miroslav Pajic, and George J. Pappas. Stochastic game approach for replay attack detection. *52nd IEEE Conference on Decision and Control*, pages 1854 – 1859, December 2013.
- [26] C.K. Gharban and Brian John Cory. Non-linear dynamic power system state estimation. *IEEE Transactions on Power Systems*, 1(3):276–283, 1986.
- [27] Branden Ghena, William Beyer, Allen Hillaker, Jonathan Pevarnek, and Alex J. Halderman. Green lights forever: analyzing the security of traffic infrastructure. *Proceedings of the 8th USENIX workshop on offensive technologies*, August 2014.
- [28] Virgil D. Gligor. A note on denial-of-service in operating systems. *IEEE Transactions on Software Engineering*, 10(3):320 – 324, May 1984.
- [29] Google. Google project wing. http://www.theatlantic.com/technology/archive/2014/08/inside-googles-secret-drone-delivery-program/379306/?single_page=true. Accessed: 2014-08-28.
- [30] Tomasz T. Gorecki, Luca Fabietti, Faran A. Qureshi, and Colin N. Jones. Experimental Demonstration of Buildings Providing Frequency Regulation Services in the Swiss Market. *Energy and Buildings*, 144:229–240, 2017.
- [31] Siddharth Goyal and Prabir Barooah. A Method for Model-Reduction of Nonlinear Thermal Dynamics of Multi-Zone Buildings. *Energy and Buildings*, 47:332–340, 2012.
- [32] Siddharth Goyal, Herbert A. Ingley, and Prabir Barooah. Occupancy-Based Zone-Climate Control for Energy-Efficient Buildings: Complexity vs. Performance. *Applied Energy*, 106:209–221, 2013.
- [33] Assane Gueye, Vladimir Marbukh, and Jean C. Walrand. Towards a metric for communication network vulnerability to attacks: A game theoretic approach. *3rd International ICST Conference on Game Theory for Networks*, May 2012.

- [34] Abhishek Gupta, Cedric Langbort, and Tamer Basar. Optimal control in the presence of an intelligent jammer with limited actions. *49th IEEE Conference on Decision and Control*, pages 1096 – 1101, December 2010.
- [35] Shirley J. Hansen and H.E. Burroughs. *Managing Indoor Air Quality*. Lulu Press, Inc., 2013.
- [36] H. Hao, A. Kowli, Y. Lin, P. Barroah, and S. Meyn. Ancillary Service for the Grid via Control of Commercial Building HVAC Systems. *American Control Conference*, June 2013.
- [37] He Hao, Yashen Lin, Anupama S. Kowli, Prabir Barooah, and Sean Meyn. Ancillary Service to the Grid Through Control of Fans in Commercial Building HVAC Systems. *IEEE Transactions on Smart Grid*, 5(4):2066–2074, 2014.
- [38] He Hao, Tim Middelkoop, Prabir Barooah, and Sean Meyn. How Demand Response from Commercial Buildings will Provide the Regulation Needs of the Grid. *the 50th Annual Allerton Conference on Communication, Control and Computing*, 2012.
- [39] Wolfgang Härdle, Hua Liang, and Jiti Gao. *Partially Linear Models*. Springer, 2000.
- [40] David Hayden, Young Hwan Chang, Jorge Goncalves, and Claire Tomlin. Sparse network identifiability via compressed sensing. *Automatica*, 68:9–17, 2016.
- [41] Julien Hendrickx, Karl H. Johansson, Raphawl Jungers, Henrik Sanberg, and Kin Cheong Sou. Efficient computations of a security index for false data attacks in power networks. *IEEE Transactions on Automatic Control*, 59(12):3194–3208, 2014.
- [42] Q. Hu, F. Oldewurtel, M. Balandat, E. Vrettos, D. Zhou, and C.J. Tomlin. Model Identification of Commercial Building HVAC Systems During Regular Operation - Empirical Results and Challenges. *American Control Conference*, 2016.
- [43] Qie Hu, Young Hwan Chang, and Claire Tomlin. Secure estimation for unmanned aerial vehicles against adversarial cyber attacks. *The 30th Congress of the International Council of the Aeronautical Sciences*, 2016.
- [44] Qie Hu, Dariush Fooladivanda, Young Hwan Chang, and Claire Tomlin. Secure state estimation and control for cyber security of the nonlinear power systems. *IEEE Transactions on Control of Network Systems*, PP(99), 2017.

- [45] Zhenyu Huang, Kevin Schneider, and Jarek Nieplocha. Feasibility studies of applying kalman filter techniques to power system dynamic state estimation. *Proceedings of the 8th International Power Engineering Conference*, 2007.
- [46] Amit Jain and N.R. Shivakumar. Impact of pmu in dynamic state estimation of power systems. *Proceedings of the 40th North American Power Symposium*, 2008.
- [47] Johan Löfberg. YALMIP: a toolbox for modeling and optimization in MATLAB. *IEEE International Symposium on Computer Aided Control Systems Design*, pages 284–289, 2004.
- [48] Kyoung-Dae Kim and Panganamala R Kumar. Cyber-physical systems: A perspective at the centennial. *Proceedings of the IEEE*, 100(Special Centennial Issue):1287–1308, 2012.
- [49] S. Koehler and Francesco Borrelli. Building Temperature Distributed Control via Explicit MPC and “Trim and Respond” Methods. *European Control Conference*, 2013.
- [50] Oliver Kosut, Liyan Jia, Robert Thomas, and Lang Tong. Malicious data attacks on the smart grid. *IEEE Transactions on Smart Grid*, 2(4):645–658, 2011.
- [51] Prabha Kundur. *Power System Stability and Control*. McGraw-Hill, 1994.
- [52] Cheolhyeon Kwon and Inseok Hwang. Hybrid robust controller design: Cyber attack attenuation for cyber-physical systems. *52nd IEEE Conference on Decision and Control*, 2013.
- [53] Cheolhyeon Kwon, Weiyi Liu, and Inseok Hwang. Security analysis for cyber-physical systems against stealthy deception attacks. *American Control Conference*, 2013.
- [54] Yashen Lin, Prabir Barooah, and Sean Meyn. Low Frequency Power Grid Ancillary Services from Commercial Building HVAC Systems. *IEEE Smart-GridComm Symposium*, 2013.
- [55] Yashen Lin, Prabir Barooah, Sean Meyn, and Timothy Middelkoop. Experimental Evaluation of Frequency Regulation From Commercial Building HVAC Systems. *IEEE Transactions on Smart Grid*, 6(2), 2015.
- [56] Shan Liu, Bo Chen, Takis Zourntos, Deepa Kundur, and Karen Butler-Purpy. A coordinated multi-switch attack for cascading failures in smart grid. *IEEE Transactions on Smart Grid*, 5(3):1183–1195, 2014.

- [57] Yao Liu, Peng Ning, and Michael K. Reiter. False data injection attacks against state estimation in electric power grids. *ACM Transactions on Information and System Security*, 14(13), May 2011.
- [58] Yilu Liu. A us-wide power systems frequency monitoring network. *IEEE Power Engineering Society General Meeting*, 2006.
- [59] The Dark Sky Company LLC. The Dark Sky Database. www.darksky.net, 2017.
- [60] Yudong Ma, Garrett Anderson, and Francesco Borrelli. A Distributed Predictive Control Approach to Building Temperature Regulation. *American Control Conference*, pages 2089–2094, 2011.
- [61] Mehdi Maasoumy, M. Razmara, M. Shahbakhti, and Alberto Sangiovanni-Vincentelli. Handling Model Uncertainty in Model Predictive Control for Energy Efficient Buildings. *Energy and Buildings*, 2014.
- [62] Mehdi Maasoumy, Catherine Rosenberg, Alberto Sangiovanni-Vincentelli, and Duncan S. Callaway. Model Predictive Control approach to Online Computation of Demand-Side Flexibility of Commercial Buildings HVAC Systems for Supply Following. *American Control Conference*, pages 1082–1089, 2014.
- [63] Jason MacDonald, Sila Kilicote, Jim Boch, Johathan Chen, and Robert Nawy. Commercial Building Loads Providing Ancillary Services in PJM. *ACEEE Summer Study on Energy Efficiency in Buildings*, 2014.
- [64] Kebina Manandhar, Xiaojun Cao, Fei Hu, and Yao Liu. Combating false data injection attacks in smart grid using kalman filter. *International Conference on Computing, Networking and Communications*, pages 16–20, February 2014.
- [65] Mohammad Hossein Manshaei, Quanyan Zhu, Tansu Alpcan, Tamer Basar, and Jean-Pierre Hubaux. Game theory meets network security and privacy. *ACM Computing Surveys*, 45(3), June 2013.
- [66] A.R. Messina, Vijay Vittal, Daniel Ruiz-Vega, and Gilberto Enrique-Harper. Interpretation and visualization of wide-area pmu measurements using hilbert analysis. *IEEE Transactions on Power Systems*, 21(4):1760–1771, 2006.
- [67] John W. Mitchell and James E. Braun. *Principles of Heating, Ventilation and Air Conditioning in Buildings*. Wiley, 2012.
- [68] Yilin Mo and Bruno Sinopoli. False data injection attacks in control systems. *Preprints of the 1st Workshop on Secure Control Systems*, 2010.

- [69] North American Electric Reliability Corporation (NERC). Critical Infrastructure Protection. <http://www.nerc.com/pa/Stand/Pages/CIPStandards.aspx>, 2017.
- [70] California Department of Water Resources. California Irrigation Management Information System (CIMIS) Station Reports. <http://www.cimis.water.ca.gov/>, 2015.
- [71] Katsuhiko Ogata. *Modern Control Engineering*. Prentice Hall, 2010.
- [72] Forward Market Operations. *PJM Manual 11: Energy and Ancillary Services Market Operations*. PJM, 83 edition, July 2016.
- [73] Miroslav Pajic, James Weimer, Nicola Bezzo, Paulo Tabuada, Oleg Sokolsky, Insup Lee, and George Pappas. Robustness of attack-resilient state estimators. *ACM/IEEE International Conference on Cyber-Physical Systems (IC-CPS)*, 2014.
- [74] Bikash Pal and Balarko Chaudhuri. *Robust Control in Power Systems*. Springer, 2006.
- [75] Alessandra Parisio, Luca Fabietti, Marco Molinari, Damiano Varagnolo, and Karl H. Johansson. Control of HVAC Systems via Scenario-Based Explicit MPC. *IEEE Conference on Decision and Control*, 2014.
- [76] Fabio Pasqualetti, Florian Dorfler, and Francesco Bullo. Cyber-physical attacks in power networks: models, fundamental limitations and monitor design. *50th IEEE conference on decision and control and european control conference*, pages 2195 – 2201, December 2011.
- [77] Luis Pérez-Lombard, José Ortiz, and Christine Pout. A Review on Buildings Energy Consumption Information. *Energy and Buildings*, 40(3):394–398, 2008.
- [78] Chris Piling. *PJM Manual 12: Balancing Operations*. PJM, 34 edition, April 2016.
- [79] Kuang rong Shih and Shyh-Jier Huang. Application of a robust algorithm for dynamic state estimation of a power system. *IEEE Transactions on Power Systems*, 17(1):141–147, 2002.
- [80] Sankardas Roy, Charles Ellis, Sajjan Shiva, Dipankar Dasgupta, Vive Shandilya, and Qishi Wu. A survey of game theory as applied to network security. *43rd Hawaii International Conference on System Sciences*, 2010.
- [81] David Ruppert, M.P. Wand, and R.J. Carroll. *Semiparametric Regression*. Cambridge University Press, 2003.

- [82] PJM Ancillary Services. Ancillary Services. <http://www.pjm.com/markets-and-operations/>, 2017.
- [83] Yasser Shoukry, Michelle Chong, Masashi Wakaiki, Pierluigi Nuzzo, Alberto L Sangiovanni-Vincentelli, Sanjit A Seshia, João P Hespanha, and Paulo Tabuada. SMT-based observer design for cyber-physical systems under sensor attacks. *International Conference on Cyber-Physical Systems*, 2016.
- [84] Yasser Shoukry, Pierluigi Nuzzo, Nicola Bezzo, Alberto Sangiovanni-Vincentelli, Sanjit Seshia, and Paulo Tabuada. Secure state reconstruction in differentially flat systems under sensor attacks. *IEEE 54th Annual Conference on Decision and Control*, 2015.
- [85] Jan Široky, Frauke Oldewurtel, Jirí Cigler, and Samuel Prívera. Experimental Analysis of Model Predictive Control for an Energy Efficient Building Heating System. *Applied Energy*, 88:3079–3087, 2011.
- [86] Joshua S. Stein. The Photovoltaic Performance Modeling Collaborative (PVPMC). *38th IEEE Photovoltaic Specialists Conference (PVSC)*, 2012.
- [87] D. Sturzenegger, D. Gyalistras, M. Morari, and R.S. Smith. Semi-Automated Modular Modeling of Buildings for Model Predictive Control. *BuildSys 2012 – Workshop of SCM SenSys Conference*, 2012.
- [88] B. Sun, P.B. Luh, Q.S. Jia, Z. Jiang, F. Wang, and C. Song. Building Energy Management: Integrated Control of Active and Passive Heating, Cooling, Lighting, Shading, and Ventilation Systems. *IEEE Transactions on Automation Science and Engineering*, 2012.
- [89] Andre Teixeira, Saurabh Amin, Henrik Sandberg, Karl H. Johansson, and Shankar Sastry. Cyber security analysis of state estimators in electric power systems. *49th IEEE Conference on Decision and Control*, pages 5991 – 5998, December 2010.
- [90] U.E. Energy Information Administration. Commercial Buildings Energy Consumption Survey (CBECS). <http://www.eia.doe.gov/emeu/cbecs/cbecs2003/overview1.html>, 2012.
- [91] US Energy Information Administration. The Annual Energy Outlook 2017. Technical report, US Energy Information Administration, January 2017.
- [92] Evangelos Vrettos, Emre C. Kara, Jason MacDonald, Goran Andersson, and Duncan Callaway. Experimental Demonstration of Frequency Regulation by Commercial Buildings - Part II: Results and Performance Evaluation. *IEEE Transactions on Smart Grid*, 2016.

- [93] Evangelos Vrettos, Emre C. Kara, Jason MacDonald, Goran Andersson, and Duncan S. Callaway. Experimental Demonstration of Frequency Regulation by Commercial Buildings - Part I: Modeling and Hierarchical Control Design. *IEEE Transactions on Smart Grid*, 2016.
- [94] Evangelos Vrettos, Frauke Oldewurtel, Fengtian Zhu, and Goran Andersson. Robust Provision of Frequency Reserves by Office Building Aggregations. *Proceedings of the 19th IFAC World Congress*, pages 12068 – 12073, 2014.
- [95] Shaobu Wang, Wenzhong Gao, and A.P. Sakis Meliopoulos. An alternative method for power system dynamic state estimation based on unscented transform. *IEEE Transactions on Power Systems*, 27(2):942–950, 2012.
- [96] Mohamed Ramadan Younis and Reza Iravani. Wide-area damping control for inter-area oscillations: A comprehensive review. *Electrical Power and Energy Conference (EPEC)*, 2013.
- [97] Jie Zhao, Khee Poh Lam, and B. Erik Ydstie. EnergyPlus model-based predictive control (EPMPC) by using MATLAB/SIMULINK and MLE+. *Proceedings of 13th Conference of International Building Performance Simulation Association*, 2013.
- [98] Liang Zhao and Ali Abur. Multi area state estimation using synchronized phasor measurement. *IEEE Transactions on Power Systems*, 20(2):611–617, 2005.
- [99] Peng Zhao, Gregor P. Henze, Sandro Plamp, and Vincent J. Cushing. Evaluation of Commercial Building HVAC Systems as Frequency Regulation Providers. *Energy and Buildings*, 67:225–235, 2013.
- [100] Datong Zhou, Qie Hu, and Claire Tomlin. Quantitative Comparison of Data-Driven and Physics-Based Models for Commercial Building HVAC Systems. *American Control Conference*, 2017.

Appendix A

Proof of Theorem 1

In the following lemmas and proposition, we assume the following:

1. $A \in \mathbb{R}^{n \times n}$ has n distinct positive eigenvalues such that $0 < \lambda_1 < \lambda_2 < \dots < \lambda_n$.
2. $C \in \mathbb{R}^{p \times n}$ is full rank.
3. The pair (A, C) is observable.
4. $\forall v_i \in \mathbb{R}^n$ where $Av_i = \lambda_i v_i$ (i.e., v_i is an eigenvector of A), $|\text{supp}(Cv_i)| > 2q$.

Recall that we want to show that given $|\text{supp}(\Phi v_i)| > 2q \cdot T$ for all eigenvectors v_i of A , then $|\text{supp}(\Phi z)| > 2q \cdot T$ for all $z \in \mathbb{R}^n \setminus \{0\}$. Now, A has n distinct eigenvalues, hence the eigenvectors of A form a basis for \mathbb{R}^n , and any $z \in \mathbb{R}^n$ can be expressed in the eigenbasis of A , i.e., $z = \sum_{i=1}^n \alpha_i v_i$. Therefore, $\Phi z = \sum_{i=1}^n \alpha_i \Phi v_i$, and thus, the only way that the number of nonzero terms in Φz may be less than $2q \cdot T$ is if too many nonzero terms in Φv_i are cancelled during this summation. In other words, there are rows that are nonzero in the vectors Φv_i , however after the scaling by α_i 's and summation, these rows become zero in Φz . Therefore our goal is to prove upper bounds on the number of such cancellations, and use these upper bounds to derive a value of T such that even in the worst case (i.e., with most number of cancellations), the number of nonzero terms in Φz is greater than $2q \cdot T$ for all $z \in \mathbb{R}^n \setminus \{0\}$.

Before we present the proof, we define the following notations: c_i^\top is the i -th row of C , $\mathbf{v} \triangleq [v_1 \ v_2 \ \dots \ v_n] \in \mathbb{R}^{n \times n}$, $\Lambda \triangleq \text{diag}\{\lambda_1, \dots, \lambda_n\} \in \mathbb{R}^{n \times n}$. For all $z \in \mathbb{R}^n$, $z = \mathbf{v} \cdot \alpha$ where $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_n]^\top$. In addition, with these notations, $\Phi z = [C\mathbf{v}\alpha; C\mathbf{v}\Lambda\alpha; \dots; C\mathbf{v}\Lambda^{T-1}\alpha]$.

We will first consider two simple cases where z is spanned by two and three eigenvectors ($m = 2$ and $m = 3$) respectively, and then generalize the results to $3 < m \leq n$. The proofs use the following result.

Lemma 3 For m real numbers $0 < \lambda_1 < \lambda_2 < \dots < \lambda_m$, and m positive integers $0 < x_1 < x_2 < \dots < x_m$, the Generalized Vandermonde matrix $GV(\lambda_1, \dots, \lambda_m; x_1, \dots, x_m)$ defined as

$$GV(\lambda_1, \dots, \lambda_m; x_1, \dots, x_m) = \begin{bmatrix} \lambda_1^{x_1} & \lambda_2^{x_1} & \dots & \lambda_m^{x_1} \\ \lambda_1^{x_2} & \lambda_2^{x_2} & \dots & \lambda_m^{x_2} \\ \vdots & \vdots & \dots & \vdots \\ \lambda_1^{x_m} & \lambda_2^{x_m} & \dots & \lambda_m^{x_m} \end{bmatrix}, \quad (\text{A.1})$$

is nonsingular.

Proof: $GV(\lambda_1, \dots, \lambda_m; x_1, \dots, x_m)$ is a submatrix of a Vandermonde matrix $V(\lambda_1, \dots, \lambda_{T-1})$ defined as

$$V(\lambda_1, \dots, \lambda_{T-1}) = \begin{bmatrix} 1 & 1 & \dots & 1 \\ \lambda_1 & \lambda_2 & \dots & \lambda_{T-1} \\ \lambda_1^2 & \lambda_2^2 & \dots & \lambda_{T-1}^2 \\ \vdots & \vdots & \dots & \vdots \\ \lambda_1^{T-1} & \lambda_2^{T-1} & \dots & \lambda_{T-1}^{T-1} \end{bmatrix}. \quad (\text{A.2})$$

where T is a positive integer, $0 < \lambda_1 < \lambda_2 < \dots < \lambda_{T-1}$, $\{x_1, x_2, \dots, x_m\} \subseteq \{0, 1, \dots, T-1\}$ and $\{\lambda_1, \dots, \lambda_m\} \subseteq \{\lambda_1, \dots, \lambda_{T-1}\}$.

$V(\lambda_1, \dots, \lambda_{T-1})$ is a Totally Positive (TP) matrix [21], and by definition of TP matrices, all minors of $V(\lambda_1, \dots, \lambda_{T-1})$, i.e., the determinant of all submatrices of $V(\lambda_1, \dots, \lambda_{T-1})$, are positive. Therefore $GV(\lambda_1, \dots, \lambda_m; x_1, \dots, x_m)$ is nonsingular. ■

Lemma 4 ($m = 2$) Consider $z = \sum_{i=1}^2 \alpha_i v_i$ (i.e., $\alpha_1 \neq 0$, $\alpha_2 \neq 0$, $\alpha_j = 0, \forall j \geq 3$):

1. If there exists $i \in \{1, 2, \dots, p\}$ such that $c_i^\top v_1 \neq 0$ and $c_i^\top v_2 \neq 0$ but the i -th row of $C\mathbf{v}\Lambda^k \alpha = 0$, i.e., there is a cancellation of the i -th row at time k , then the i -th row of $C\mathbf{v}\Lambda^l \alpha \neq 0$ for all $l \in \{0, \dots, T-1\}$ where $l \neq k$. In other words, there cannot be another cancellation of the i -th row at any other time step, or equivalently, there is a maximum of one cancellation of the i -th row over T time steps.

2. If we choose $T > \frac{\min\{s_1, s_2\}}{\max\{s_1, s_2\} - 2q}$, then $|\text{supp}(\Phi z)| > 2q \cdot T$.

($m = 3$) Consider $z = \sum_{i=1}^3 \alpha_i v_i$ where $\alpha_1 \neq 0$, $\alpha_2 \neq 0$, $\alpha_3 \neq 0$ and $\alpha_i = 0$ for $i = \{4, 5, \dots, n\}$.

3. $\sum_{k=0}^{T-1} s_{r,123}^k \leq 2 \cdot s_{123}$ where $s_{123} = |L_{123}| = |\text{supp}(Cv_1) \cap \text{supp}(Cv_2) \cap \text{supp}(Cv_3)|$.

4. If we choose $T > \frac{p + \min\{s_1, s_2, s_3\}}{\max\{s_1, s_2, s_3\} - 2q}$, then $|\text{supp}(\Phi z)| > 2q \cdot T$.

Proof: (1) (Suppose not) There exist $l \neq k$ and $l \in \{0, \dots, T-1\}$ such that $c_i^\top \mathbf{v} \Lambda^l \alpha = c_i^\top \mathbf{v} \Lambda^k \alpha = 0$:

$$\begin{aligned} 0 &= c_i^\top \begin{bmatrix} v_1 & v_2 \end{bmatrix} \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix} \begin{bmatrix} \lambda_1^l \\ \lambda_2^l \end{bmatrix} \\ &= c_i^\top \begin{bmatrix} v_1 & v_2 \end{bmatrix} \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix} \begin{bmatrix} \lambda_1^k \\ \lambda_2^k \end{bmatrix} \end{aligned}$$

Without loss of generality, assume $l < k$. We can reformulate above equation as follows:

$$\begin{bmatrix} \lambda_1^l & \lambda_2^l \\ \lambda_1^k & \lambda_2^k \end{bmatrix} \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix} \begin{bmatrix} v_1^\top \\ v_2^\top \end{bmatrix} c_i = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Now, the first matrix on the left hand side (LHS) is a Generalized Vandermonde matrix $GV(\lambda_1, \lambda_2; l, k)$ with $0 < \lambda_1 < \lambda_2$ and $0 < l < k$, thus by Lemma 3 it is nonsingular. The second matrix on the LHS is also nonsingular as $\alpha_1 \neq 0$ and $\alpha_2 \neq 0$. Therefore we must have $v_1^\top c_i = v_2^\top c_i = 0$ (contradiction, since $c_i^\top v_1 \neq 0$ and $c_i^\top v_2 \neq 0$ by assumption).

(2) Let L_1, L_2, L_{12} be three disjoint subsets of $\{1, \dots, p\}$ such that $L_1 = \text{supp}(Cv_1) \cap \text{supp}(Cv_2)^c$, $L_2 = \text{supp}(Cv_2) \cap \text{supp}(Cv_1)^c$, and $L_{12} = \text{supp}(Cv_1) \cap \text{supp}(Cv_2)$ where the superscript c represents the set complement. Then, $\text{supp}(Cv_1) = L_1 \oplus L_{12}$, $\text{supp}(Cv_2) = L_2 \oplus L_{12}$, $s_1 \triangleq |\text{supp}(Cv_1)| = |L_1 \oplus L_{12}| > 2q$, $s_2 \triangleq |\text{supp}(Cv_2)| = |L_2 \oplus L_{12}| > 2q$ and $s_{12} \triangleq |L_{12}| \leq \min\{s_1, s_2\}$. Also, possible cancellations only occur in the subset L_{12} by definition.

$$\begin{aligned} |\text{supp}(\Phi z)| &= |\text{supp}(\alpha_1 \Phi v_1 + \alpha_2 \Phi v_2)| \\ &= T \cdot (s_1 - s_{12}) + T \cdot (s_2 - s_{12}) \\ &\quad + \sum_{k=0}^{T-1} (s_{12} - s_{r,12}^k) \\ &= T \cdot (s_1 + s_2 - s_{12}) - \sum_{k=0}^{T-1} s_{r,12}^k \\ &\geq T \cdot (s_1 + s_2 - \min\{s_1, s_2\}) - \sum_{k=0}^{T-1} s_{r,12}^k \end{aligned}$$

where $s_{r,12}^k$ is the number of cancelled support in L_{12} at time step k . More specifically,

$i \in s_{r,12}^k$ if $c_i^\top v_1 \neq 0$ and $c_i^\top v_2 \neq 0$, but $c_i^\top \mathbf{v} \Lambda^k \alpha = 0$. From (1), we have the followings:

$$\begin{aligned} s_{r,12}^0 &\leq |L_{12}| \\ s_{r,12}^1 &\leq |L_{12}| - s_{r,12}^0 \\ s_{r,12}^2 &\leq |L_{12}| - s_{r,12}^0 - s_{r,12}^1 \\ &\vdots \\ s_{r,12}^{T-1} &\leq |L_{12}| - \sum_{k=0}^{T-2} s_{r,12}^k \end{aligned}$$

Thus, $\sum_{k=0}^{T-1} s_{r,12}^k \leq |L_{12}| \leq \min\{s_1, s_2\}$, and

$$|\text{supp}(\Phi z)| \geq T \cdot \max\{s_1, s_2\} - \min\{s_1, s_2\} > 2q \cdot T.$$

(3) Claim: the i -th row of $C\mathbf{v}\Lambda^k\alpha$ is cancelled at most 2 times over T time steps, i.e., for at most 2 distinct values of $k \in \{0, 1, \dots, T-1\}$.

(Suppose not) There exist 3 distinct time steps $d, e, f \in \{0, 1, \dots, T-1\}$ such that $c_i^\top \mathbf{v} \Lambda^d \alpha = c_i^\top \mathbf{v} \Lambda^e \alpha = c_i^\top \mathbf{v} \Lambda^f \alpha = 0$. Without loss of generality, assume $d < e < f$:

$$\begin{bmatrix} \lambda_1^d & \lambda_2^d & \lambda_3^d \\ \lambda_1^e & \lambda_2^e & \lambda_3^e \\ \lambda_1^f & \lambda_2^f & \lambda_3^f \end{bmatrix} \begin{bmatrix} \alpha_1 & 0 & 0 \\ 0 & \alpha_2 & 0 \\ 0 & 0 & \alpha_3 \end{bmatrix} \mathbf{v}^\top c_i = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Again, the first matrix on the LHS is a Generalized Vandermonde matrix satisfying the conditions of Lemma 3, hence it is nonsingular. The second matrix on the LHS is also nonsingular as $\alpha_1 \neq 0$, $\alpha_2 \neq 0$ and $\alpha_3 \neq 0$. Therefore we must have $\mathbf{v}^\top c_i = 0$ (contradiction). A similar derivation as in Lemma 4 then shows $\sum_{k=0}^{T-1} s_{r,123}^k \leq 2 \cdot s_{123}$.

(4) Consider Cv_1, Cv_2, Cv_3 and Φz :

$\text{supp}(Cv_1)$	$\text{supp}(Cv_2)$	$\text{supp}(Cv_3)$
L_1	$\mathbf{0}$	$\mathbf{0}$
$\mathbf{0}$	L_2	$\mathbf{0}$
$\mathbf{0}$	$\mathbf{0}$	L_3
L_{12}	L_{12}	$\mathbf{0}$
$\mathbf{0}$	L_{23}	L_{23}
L_{13}	$\mathbf{0}$	L_{13}
L_{123}	L_{123}	L_{123}

Without loss of generality, assume $s_3 \geq s_2 \geq s_1$ (recall $s_1 = |L_1 \oplus L_{12} \oplus L_{13} \oplus L_{123}| = |L_1| + s_{12} + s_{13} + s_{123}$):

$$\begin{aligned}
|\text{supp}(\Phi z)| &= T \cdot (s_1 - s_{12} - s_{13} - s_{123}) \\
&\quad + T \cdot (s_2 - s_{12} - s_{23} - s_{123}) \\
&\quad + T \cdot (s_3 - s_{23} - s_{13} - s_{123}) \\
&\quad + \sum_{k=0}^{T-1} \left((s_{12} - s_{r,12}^k) + (s_{23} - s_{r,23}^k) \right. \\
&\quad \quad \left. + (s_{13} - s_{r,13}^k) + (s_{123} - s_{r,123}^k) \right) \\
&= T \cdot (s_3 + s_1 + s_2 - s_{12} - s_{13} - s_{23} - 2 \cdot s_{123}) \\
&\quad - \sum_{k=0}^{T-1} (s_{r,12}^k + s_{r,23}^k + s_{r,13}^k + s_{r,123}^k) \\
&\geq T \cdot s_3 - p - s_{123} \geq T \cdot s_3 - p - \min\{s_1, s_2, s_3\} \\
&> 2q \cdot T
\end{aligned}$$

where $\sum_{k=0}^{T-1} (s_{r,12}^k + s_{r,23}^k + s_{r,13}^k) \leq s_{12} + s_{23} + s_{13} \leq p - s_{123}$, $\sum_{k=0}^{T-1} s_{r,123}^k \leq 2 \cdot s_{123}$ and $s_{123} \leq \min\{s_1, s_2, s_3\}$ and note that $T > \frac{p + \min\{s_1, s_2, s_3\}}{\max\{s_1, s_2, s_3\} - 2q}$. ■

Proposition 2 Consider m eigenvector combinations ($n \geq m \geq 2$). Then, the total number of cancellations over T time steps satisfies $\sum_{k=0}^{T-1} s_{r,12\dots m}^k \leq (m-1) \cdot s_{12\dots m}$ where $s_{12\dots m} = |\text{supp}(Cv_1) \cap \dots \cap \text{supp}(Cv_m)|$.

Proof: Claim: the i -th row of $Cv\Lambda^k\alpha$ is cancelled at most $(m-1)$ times over T time steps, i.e., for at most $(m-1)$ distinct values of $k \in \{0, 1, \dots, T-1\}$ (Suppose not)

$$\begin{bmatrix} \lambda_1^d & \lambda_2^d & \dots & \lambda_m^d \\ \lambda_1^e & \lambda_2^e & \dots & \lambda_m^e \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1^r & \lambda_2^r & \dots & \lambda_m^r \end{bmatrix} \begin{bmatrix} \alpha_1 & 0 & \dots & 0 \\ 0 & \alpha_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \alpha_m \end{bmatrix} \mathbf{v}^\top c_i = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Again, the first matrix on the LHS is a Generalized Vandermonde matrix satisfying the conditions of Lemma 3, hence it is nonsingular. The second matrix on the LHS is also nonsingular as $\alpha_i \neq 0$ for all $i \in \{1, 2, \dots, m\}$. Therefore we must have $\mathbf{v}^\top c_i = 0$ (contradiction). A similar derivation as in Lemma 4 then shows $\sum_{k=0}^{T-1} s_{r,12\dots m}^k \leq (m-1) \cdot s_{12\dots m}$. ■