Secure Detection in Cyberphysical Control Systems

Submitted in partial fulfillment of the requirements for

the degree of

Doctor of Philosophy

in

Electrical and Computer Engineering

Rohan Chabukswar

B.Tech., Engineering Physics, Indian Institute of Technology Bombay
M.S., Electrical and Computer Engineering, Carnegie Mellon University

Carnegie Mellon University
Pittsburgh, PA

May, 2014

To my parents, who tolerated me being in school for 26 out of my 28 years of my life. I finally have one degree more than either of you, although I will never quite fit the role of “Dr. Chabukswar”. That will always be you, Dad.

Acknowledgements

None of the work presented herein would have been possible without the support of numerous people that surrounded me and sheltered me (some would say unsuccessfully) from becoming insane.

First and foremost, I would like to thank Bruno Sinopoli for his invaluable guidance, ceaseless encouragement, continuous support, and infinite patience. I could not have asked for a more perfect advisor, and I greatly value his advice — academic, sartorial, and life.

I thank my thesis committee members, Prof. Anthony Rowe and Prof. Pulkit Grover of Carnegie Mellon University, and Prof. Henrik Sandberg of the Royal Institute of Technology (Kungliga Tekniska Högskolan), Stockholm, for their insight and expert opinions.

For my friends who have occupied offices, past and present, in the Porter Hall B Level workspace — Luca Parolini and Ajinkya Bhave, thanks for being partners-in-crime around town. I thank June Zhang, for all the madness around the office, Kyri Baker for being one of the coolest persons I know, and Jim Weimer, for giving me a level of craziness to aspire to. Aurora Schmidt, Nikos Arechiga, Akshay Rajhans, Joel Harley, Kyle Anderson, JY Joo, Sergio Pequito, Javad Mohammadi, Andrew Hsu, and Joya Deri, thanks for all the inane conversations and arguments with me and June around lunch time. I will miss these dearly, although I still hope to join in on them sometimes. Matthias Althoff, Jonathan Donadee, Milos Cvetkovic, and Evgeny Toropov, thanks for getting me to join in fun stuff around Pittsburgh. Anit Sahu,

Steven Aday, Nipun Popli, Subhro Das, and everyone else, thank you for all the celebrations, all the parties, all the laughs, all the madness and scandals around the office. I would like to thank my other friends in and around CMU, especially the Quiz Club, which kept me replete with inconsequential knowledge, and my friends around Pittsburgh, Shishir, Sunny, Vinay et al, for the dinners and movies on Friday nights.

I would like to thank the other people in Bruno's group for working with me — Yilin Mo, Dragana Bajovic, Sabina Zejnilovic, Niranjini Rajagopal, Xiaoqi Yin, Sean Weerrakody, and Xiaofei Liu.

I would like to thank Claire, Carol, and other people around the department for working behind the scenes in making everything work well, for keeping us supplied with coffee, and for patiently deciphering the ball of reimbursement receipts that I used to bring back from my domestic and international travels.

I would like to take this opportunity to thank Radhika Marathe, one of my oldest and closest friends, who for years has been listening to me rant about academic life, professional life, social life, my parents, and other countless topics. I would like to thank Devaki Erande, who has been a source of constant support, even at the lowest points in my life, most of which were during my PhD years. To my other friends outside Pittsburgh, Aniket, Robin, Shriharsh, Sangram, Kartik, Rahul, Meenakshi, you guys have always been around for me, and if I have been remiss in losing contact with some of you for periods of time because of grad school, I promise I will get back in touch.

On a rather unconventional note, I would like to thank Tony Horton, who, unbeknownst to him, helped me regain my physical and mental health with his P90X workouts (*mens sana in corpore sano*, after all), and drove home several maxims including “No Excuses”.

I would like to thank my whole family for being a part of who I am today.

But most of all, my heartfelt gratitude goes to my mom and dad, who were always there for me, who encouraged me without pushing, inspired me without forcing, supported me without directing, and guided me without steering. You always approved of me doing what I wanted, and prompted me to do my best in everything I took up. It has been a long journey through my schooling years, and I hope I will make you proud and attain the heights which you envisage for me.

Acknowledgements

This research was supported in part by CyLab at Carnegie Mellon by grant DAAD19-02-1-0389 from the Army Research Office, grant NGIT2009100109 from the Northrop Grumman Information Technology, Inc. Cybersecurity Consortium, and grant 0955111 from the National Science Foundation. The views and conclusions contained here are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either express or implied, of ARO, CMU, CyLab, NSF, NGC, or the U.S. Government or any of its agencies.

Abstract

A SCADA system employing the distributed networks of sensors and actuators that interact with the physical environment is vulnerable to attacks that target the interface between the cyber and physical subsystems. An attack that hijacks the sensors in an attempt to provide false readings to the controller (for example, the Stuxnet worm that targeted Iran's nuclear centrifuges) can be used to feign normal system operation for the control system, while the attacker can hijack the actuators to send the system beyond its safety range. This thesis extends the results of a previously proposed method. The original method proposed addition of a randomized "watermarking" signal and checking for the presence of this signal and its effects in the received sensor measurements. Since the control inputs traverse the cyberphysical boundary and make their effects apparent in the sensor measurements, they are employed to carry this watermarking signal through to the system and back to the SCADA controller. The sensor measurements are compared to the expected measurements (calculated using a suitably delayed model of the system within the controller). This methodology is based on using the statistics of the linear system and its controller. The inclusion of a randomized signal on the control inputs induces an increase in the performance cost of the physical system. This thesis proposes a method of optimization of the watermarking signal based on the trade-off between this performance cost and the attack detection rate, by leveraging the distribution the watermarking signal over multiple inputs and multiple outputs. It is further proved that regardless of the number of inputs and outputs in the system, only one watermarking signal needs to be generated. This optimization, and its necessity in improving the effectiveness of the detector without detriment to the performance cost, are demonstrated on a simulated chemical plant. The thesis also proposes another methodology that does not rely on these statistics, but is instead based on calculating the correlation between the received sensor measurements and the expected measurements accrued from the model inside the controller.

Generalizing the form of attack even further to attacks that target the integrity of the data sent to the actuators and received from the sensors, this thesis demonstrates the effect of such integrity attack on electricity market operations, where the attacker successfully uses a vulnerability in the Global Position System to break synchronicity among dispersed phasor measurements to gain a competitive advantage over other bidders in the electricity market. In an effort to make state estimation robust against integrity attacks, the sensors and states are modeled as binary variables. Sensor networks use binary measurements and state estimations for several reasons, including communication and processing overheads. Such a state estimator is vulnerable to attackers that can hijack a subset of the sensors in an effort to change the state estimate. This thesis proposes a method for designing the estimators using the concept of invariant sets. This methodology relies on identifying the sets of measurement vectors for which no amount of manipulation by the attacker can change estimate, and maximizing the probability

that the sensor measurement vector lies in this set. Although the problem of finding the best possible invariant sets for a general set of sensors has double-exponential complexity, by using some simplifications on the types of sensors, this can be reduced significantly. For the problem that employs all sensors of the same type, this method reduces to a linear search. For sensors that can be classified into two types, this complexity reduces to a search over a two-dimensional space, which is still tractable. Further increase in the confidence of the estimate can be achieved by considering the correlation between the sensor measurements.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis examines the security of control and estimation in cyberphysical systems.

Cyberphysical systems (CPS) often employ distributed networks of embedded sensors and actuators ([1]) that interact with the physical environment, and are monitored and controlled by a Supervisory Control and Data Acquisition (SCADA) system. Distributed sensors and actuator networks are often seen in varied applications, such as critical infrastructure monitoring, autonomous vehicle control, healthcare, etc.

Given the ubiquity of cyberphysical systems, and the reliance on their performance, incentives are abundant for miscreants to attack such systems, from simple economic reasons (reducing gas bills), and advantages over industrial competitors (manipulating differential electricity pricing), to political espionage and sabotage (derail national scientific and military programs) and full-fledged terrorism (cause communications breakdown, traffic disruptions). Isolation of CPS networks and controllers from the Internet can only offer a limited amount of protection, not only because of the advent of increasingly "smart" cyberphysical systems like Smart Grids, which require Internet access, but also because of the increasing deployment of sensors to remote locations where the sensors themselves, and the communications to and from them, cannot be adequately monitored for security.

Additionally, organized criminals, industrial spies, and global terrorists have proved themselves adept at introducing malware into heavily secured and isolated networks by relying on human errors. An example of alleged digital warfare, waged against Iran's Natanz nuclear facility, is the Stuxnet worm, which seems to have been specially designed to reprogram certain industrial centrifuges and make them fail in a way that was virtually undetectable ([2]). The worm, which was chiefly used in coordination with espionage malware, was introduced by infected USB flash drives ([3]), and further used peer-to-peer calls to infect other computers inside private networks ([4]). It is evident that relying on isolation of networks and components, and in general, security with obscurity, is at best only a short-term solution. The worm itself has been claimed (by Edward Snowden, in an interview with the German newspaper "Der Spiegel", [5]) to be a part of a US-Israeli operation dubbed "Operation Olympic Games" ([6]). Other speculations and allegations have flown back and forth, accusing various national intelligence agencies and even the manufacturer, but irrespective of the attacker, target or intention, this worm has indubitably brought to light serious security susceptibilities in industrial control systems. This attack resonated with a recent concern in distributed control system security, whereby an attacker could modify the software or environment of some of the networked sensors and/or actuators, to launch a coordinated attack against the system infrastructure.

In view of the present threat of global terrorism, a power grid failure, a local breakdown of telecommunications system, or a disruption of air traffic control (ATC) at a major hub could all be executed as an antecedent to a full-fledged invasion. Such threats have been predicted ([7]) and even made into movies. CPS infrastructures like power grids, telecommunication networks, air traffic controllers — vital to the normal operation of a society — are safety critical, and a successful attack on one of them, or worse, a coordinated attack on two or more of them, can significantly hamper the economy, endanger human lives or even make the community vulnerable to foreign aggression. This makes the design of secure cyberphysical systems of paramount importance.

A conventional method of security is using symmetric and asymmetric encryption and decryption to secure the communications. While this approach might be sufficient for day-to-day usage, in cases of national security a more robust security mechanism is called for. Cryptographic keys are broken and stolen daily, but even if they were secure, an attacker could directly attack the physical environment of the components, without even touching the communication network. Such an attack is feasible when sensors and actuators are spatially distributed in remote locations. There are other methods of approaching CPS security, most of which rely either on the information content of the system (confidentiality, integrity, availability), or on the robustness of controllers and estimation, detection and identification algorithms. The problem with concentrating on the information content is the lack of a system model, which can blind the detector to a wide variety of attacks (for example, lowering electricity bills by bypassing the meter). On the other hand, robust controllers and algorithms tend to assume random, uncoordinated failures, which is hardly the case during an attack.

Consequently, system knowledge and traditional-cyber security are both essential to ensure the secure operation of safety critical cyberphysical systems.

The rest of the thesis is organized as follows: In Chapter 2, the replay attack problem is formulated and the previous theoretical work on the problem is indicated. A new countermeasure that can effectively detect the replay at the expense of control performance, for a general form of controller and detector is also introduced and analyzed theoretically. In Chapter 3, a practical problem requiring such a defense system against a replay attack is put forth. The countermeasures outlined in Chapter 2 are applied to the system in succession, and compared.

In Chapter 4, a more sophisticated problem requiring a defense against integrity attacks is presented. The attack methodologies for different levels of attacker resources are provided. The current state-of-the-art defense measures are indubitably a part of the system and the attacker is shown to circumvent them. The extent of disruption for a particular attacker objective is simulated.

In Chapter 5, a basic version on the integrity attack, using static detection, is formulated, and the previous theoretical work on the problem is indicated, along with the limitations for practical application. The rationalization for simplification of the problem in practical situation is provided. The optimal detector for the simplified case is provided. The problem is then generalized for a wider class of practical application and the optimal detector is provided for this situation. The problem formulation is further generalized, in a way that can be applied to the system of Chapter 4.

Finally, chapter 6 concludes with the results and summary of the thesis, and discusses the direction of future work.

1.1 Motivation

Cyberphysical attacks can be classied by three parameters ([8]):

1. *A priori* System Knowledge: Attacker's knowledge of system parameters,
2. Disclosure Power: Attacker's access to real-time system information, and
3. Disruption Power: Attacker's capability to disturb the system.

Figure 1.1 classifies four examples of cyberphysical attacks. An eaversdropping attack is one where the attacker needs to neither know the system parameters, nor disrupt the system — he only snoops on the sensor values. Similarly, a denial-of-service (DoS) attack is one where the attacker needs only to disrupt the system by preventing communication between the components — he does not need to know any system parameter or current state. In contrast, a replay attack requires both the current system state as well as disruption on the part of the attacker to be successful.

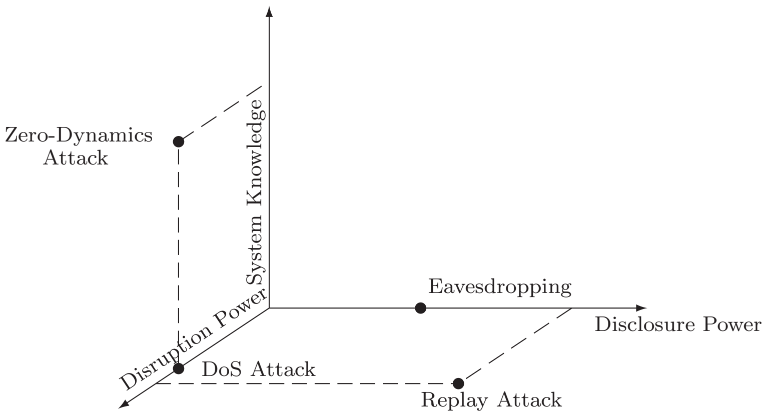


Figure 1.1: Classification of Cyberphysical Attacks

This thesis deals with integrity attacks on the sensors of a SCADA system. Sensors are usually the most vulnerable components of a secure system — they are constrained in energy,which constrains their communicating and computing capabilities. Security implemented by using sophisticated encryption techniques might be too heavy for the sensors' limited computing abilities, which makes each sensor a weak link in the chain of security. The problem is further exacerbated because the number of sensors of a system is usually large, larger than the number of actuators. Any chain is only as strong as its weakest link, and to ensure the security of safety-critical cyber-physical systems, the security of the sensors and the communication from them is of paramount importance.

The applications of resilience to integrity attacks are myriad. These methodologies can be implemented on process control systems (as evidenced by the simulations on a chemical plant), smart grids and other distribution networks, and so on.

1.2 Thesis Contributions

This work builds on the previous theoretical results. The first paper by Mo and Sinopoli ([9]) that proposed the original problem and attack strategy introduced the concept of physical watermarking, with some simulations on a model of a moving vehicle.

This thesis further enhances the original technique and makes the design process more systematic. In the case of Multi-Input-Multi-Output (MIMO) systems, a methodology to minimize the loss in control performance caused by the random watermarking signal is provided. This exploits the myriad of ways the watermarking signal can be disseminated through the multiple inputs — strength of signal on each input, interdependence of these components (or lack thereof), etc. — to extract the optimal form of the watermarking signal. It is further substantiated the this optimal watermarking signal, being the hallmark of the physical system, is independent of the specific form of the optimization problem, thus logically dividing the design of the security feature into two independent parts. The first part, which can be executed offline, involves setting up and solving a considerable linear programming problem. The form of the watermark, obtained from this part, is used in the second step, which designs the strength of the watermarking signal. This subsequent step can be executed online, and the strength of the signal can be recomputed on the fly, based on current security threat levels and the required false-alarm/detection rate values required.

Moreover, it is demonstrated that the optimal watermarking signal will always have all components dependent on one — in essence, only one random number generator is required to generate the watermark. This results caps the requirements on computational power and the extent of the trusted computational base essential to the security of the control system.

All these techniques, and the comparison of their performance, are demonstrated on a single system, a linearized version of a famous control problem of a chemical plant.

In order to demonstrate, as an example, the economic impact of attacks on cyberphysical systems, an attack on the Phasor Measurement Units (PMUs), currently being installed in smart grids, was designed. This attack was based on the previous work of Xie et al ([10]), who studied the economic impact of a potential class of integrity cyber attacks on electric power market operations. A simulation of this attack was carried out to demonstrate that, even with a very restricted attack targeting just the timing reference of such PMUs, an intelligent attacker can manipulate the locational electricity prices, with a view to maximizing profits for the bidding entities involved in the market.

The necessity of defense against such attacks, combined with the complexity of tackling a continuous-state non-linear system, gives rise to the formulation of an approximation of such a system, focusing on binary state variables and binary sensors. These sensors are then segregated into two classes in order to better resemble the smart grid, where, although the PMUs promise better measurements and state estimation, are too expensive to be installed on more than a fraction of the buses (30% penetration is the most optimistic scenario that industries are targeting). Building on a previous technique, this thesis attempts to reduce the complexity of the combinatorial formulation using simplifying assumptions.

A closer approximation to physical systems like smart grids is further achieved by considering the effect of correlation in the sensor measurements induced by the physics of the system in question. The adherence to physical laws causes the sensor readings to have significant correlations across the grid, and this correlation can be leveraged to further restrict the possibility of an undetected attack occurring.

Chapter 2

Replay Attacks: Theoretical Problem

In this chapter, first the methodology proposed by Mo and Sinopoli ([9]) for detecting replay attacks in general linear systems is briefly reviewed, after which the new system and attack models are introduced. The optimization of the existing authentication signal is then proposed, to maximize the detection rate while keeping the cost-increase bounded. The authentication signal for the new system is proposed, and its optimization is discussed.

2.1 Previous Work

The importance of addressing the security of cyberphysical systems has been stressed by the research community, by, among others, Byers and Lowe ([11], who have summarized a number of industrial security incidents, and Cárdenas et al ([12]), who first identified and defined the problem of secure control. In a later paper, Cardenas et al ([13]) discuss the cyberphysical impact of denial-of-service (DoS) attacks, which interrupt information flow from the sensors, actuators and the control system, and deception attacks that compromise the integrity of data packets. DoS attacks and a feedback control design resilient to them are further discussed by Amin et al ([14]), which concentrates on the security of the ''cyber'' aspect of the system. In contrast, this thesis assumes that the communication within the different components of the system is secure, and instead focuses on the security of the boundary between the cyber and physical aspects of the system.

A substantial amount of research has been carried out in analyzing, detecting and failure-handling CPS. Sinopoli et al study the effect of random packet drops on controller and estimator performance ([15], [16]). Several failure-detection schemes in dynamic systems are reviewed by Willsky ([17]). Some CPS scenarios, for example, those proposed by Stengel and Ray ([18]), are capable of utilizing results from robust control, where the authors concentrate on designing controllers for systems with unknown or uncertain parameters. While these works make the assumption that failures are either random or benign, a shrewd attacker, such as is considered in this thesis, can carefully construct an attack strategy to deceive detectors and make even the most robust controllers fail.

Alpcan and Başar ([19]) applied game theoretic principles formally to intrusion detection to develop a decision and control framework. Their work considers the treatment of intrusion-detection sensors, not on the actual scheme of detection that each sensor employs. Controllability and observability of linear systems has been analyzed using graph theory by Sundaram and Hadjicostis ([20]), who provide methods for reaching consensus in the presence of malicious agents. The proposed methods are combinatorial in nature and thus computationally expensive. In scenarios such as distributed sensor environments, computational cost can be prohibitive.

Robust estimation using sensors in untrusted environments has been investigated by Lazos and Poovendran ([21]), and again by Lazos et al ([22]), where the authors propose robust localization algorithms. Their work concentrates on solely on securely determining the location information of the sensors. In contradistinction, this thesis focuses on the integrity of the actual sensor data. Pasqualetti et al ([23], [24]) consider intentional malicious data attack, and address the problem of distributed monitoring and intrusion detection. Distributed formation control in the presence of attackers is studied by Zhu and Martínez ([25]) where a distributed control algorithm using online adaptation is proposed. All of these scenarios, however, unlike the present work, consider a noiseless process and environment, which is unlikely to be the case in practical applications.

Giani et al ([26]) address the problem of secure and resilient power transmission and distribution, and point out several potential threats in modern power systems. A comprehensive survey of current results in networked control systems has been carried out by Hespanha et al ([27]). Dán and Sandberg ([28]) analyze stealth attacks on power system state estimators, and use a static system formulation unlike the current work. Sandberg et al ([29]) study the analysis of large-scale power networks of using proposed security indices.

Secure state-estimation and control of systems under attack is further investigated by Fawzi et al ([30], [31]). The security of power networks, however, focus on static systems, contrary to the fundamental formulation of a Linear Time-Invariant (LTI) system analyzed in this paper.

Considerable research has been devoted to constructing estimators that are not unduly affected by outliers or other small departures from model assumptions (Maronna et al [32], Huber and Ronchetti [33]), which can be used to nullify the effect of outliers. However, the case of an attack is quite different from randomly occurring outliers, and such methods need to be reformulated for CPS. Bad data detection has been used in power grids for a long time (Abur and Expósito [34]). Liu et al ([35]) and Sandberg et al ([29]) consider how an attacker can design and inject inputs into measurements to change state estimation results.

2.2 Problem Formulation

We consider a discrete-time linear time-invariant system with n state variables. The physical part of system has p actuators as control inputs, and m sensors that measure a linear function of the system state. The cyber-part includes a communication network that communicates all the sensor readings to a base station at each discrete time step. The base station is equipped with a state estimator in the form of a Kalman filter, a linear controller that minimizes a quadratic cost, and a detector that analyzes the statistics of the noise to detect an attack.

This subsection presents the problem formulation by deriving the Kalman filter, the LQG controller and χ^2-detector for the case under study. The notation developed below is used for the remainder of the section.

2.2.1 System Dynamics

Consider a linear, time invariant (LTI) system, with the following state dynamics:

$$x_{k+1} = Ax_k + Bu_k + w_k, \tag{2.1}$$

where $x_k \in \mathbb{R}^n$ is the vector of state variables at time k, $u_k \in \mathbb{R}^p$ is the control input, $w_k \in \mathbb{R}^n$ is the process noise at time k, and x_0 is the initial state. We assume w_k, x_0 are independent Gaussian random variables, $x_0 \sim \mathcal{N}(\bar{x}_0, \Sigma)$, $w_k \sim \mathcal{N}(0, Q)$.

A sensor network monitors the system described in Equation (2.1). At each step all the sensor readings are sent to a base station. The observation equation can be written as

$$y_k = Cx_k + v_k, \tag{2.2}$$

where $y_k \in \mathbb{R}^m$ is a vector of measurements from the sensors and $v_k \sim \mathcal{N}(0, R)$ is the measurement noise independent of x_0 and w_k.

It is assumed that the system operator wants to minimize the following infinite-horizon linear quadratic Gaussian cost:

$$J = \min \lim_{T\to\infty} E\frac{1}{T}\left[\sum_{k=0}^{T-1}\left(x_k^T W x_k + u_k^T U u_k\right)\right], \tag{2.3}$$

where W, U are positive semi-definite matrices that decide the relative weight given to the deviation of the state variables from the operating point and the power required for the control inputs. u_k is measurable with respect to $y_0, y_1, \ldots, y_k$, i.e., u_k is a function of the previous observations. It is a well-known result that the separation principle holds in this case, and the optimal solution of Equation (2.3) is a combination of Kalman filter and LQG controller.

2.2.2 The Estimator — Kalman Filter

The Kalman filter provides the optimal state estimate $\hat{x}_{k|k}$ and takes the following form:

$$\begin{aligned}
\hat{x}_{0|-1} &= \bar{x}_0,\ P_{0|-1} = \Sigma, && (2.4)\\
\hat{x}_{k+1|k} &= A\hat{x}_k + Bu_k,\ P_{k+1|k} = AP_kA^T + Q, &&\\
K_k &= P_{k|k-1}C^T\left(CP_{k|k-1}C^T + R\right)^{-1}, &&\\
\hat{x}_k &= \hat{x}_{k|k-1} + K_k\left(y_k - C\hat{x}_{k|k-1}\right), && (2.5)\\
P_k &= P_{k|k-1} - K_kCP_{k|k-1}. &&
\end{aligned}$$

Although the Kalman filter uses a time varying gain K_k, it is known that this gain will converge if the system is detectable. In practice the Kalman gain usually converges in a few steps. Hence, P and K can be defined as

$$P \triangleq \lim_{k\to\infty} P_{k|k-1},\ K \triangleq PC^T \left(CPC^T + R\right)^{-1}. \tag{2.6}$$

Since control systems usually run for a long time, for all practical purposes, the system can be assumed be at steady state since the beginning. That is, the initial condition $\Sigma = P$ is assumed, which reduces the Kalman filter to a fixed gain estimator, taking the following form:

$$\begin{aligned} \hat{x}_{0|-1} = \bar{x}_0,\ \hat{x}_{k+1|k} &= A\hat{x}_k + Bu_k, \\ \hat{x}_k = \hat{x}_{k|k-1} &+ K\left(y_k - C\hat{x}_{k|k-1}\right). \end{aligned} \tag{2.7}$$

2.2.3 The Controller — Linear Quadratic Gaussian Controller

The LQG controller is a fixed gain linear controller based on the optimal state estimation $\hat{x}_k$, and takes the following form:

$$u_k = u_k^* = -\left(B^T SB + U\right)^{-1} B^T SA\hat{x}_k, \tag{2.8}$$

where u_k^* is the optimal control input and S satisfies the Riccati equation

$$S = A^T SA + W - A^T SB\left(B^T SB + U\right)^{-1} B^T SA. \tag{2.9}$$

Let $L \triangleq -\left(B^T SB + U\right)^{-1} B^T SA$, then $u_k^* = L\hat{x}_k$. The optimal value of objective function given by the optimal estimator and controller in this case is

$$J = \mathrm{tr}\,(SQ) + \mathrm{tr}\left[\left(A^T S A + W - S\right)(P - KCP)\right]. \tag{2.10}$$

2.2.4 χ^2 Failure Detector

The χ^2 detector ([36], [37]) is widely used to detect anomalies in control systems. It leverages the fact that the residues after Kalman estimation and LQG control are zero-mean Gaussian, making the weighted sum-of-squares of these residues follow a χ^2-distribution. Use of other detectors, and in fact other combinations of estimators, controllers and detectors will be commented upon in section 2.7.

Prior to introducing the detector, it is necessary to characterize the probability distribution of the residue of the Kalman filter:

Theorem 1. *For the LTI system defined in Equation* (2.1) *with the Kalman filter and the LQG controller, the residues $y_i - C\hat{x}_{i|i-1}$ of the Kalman filter are independent and identically distributed (i.i.d.) Gaussian distributed with zero mean and covariance $\mathscr{P}$, where $\mathscr{P} = CPC^T + R$.*

Proof. The proof is given by Mehra and Peschon ([36]). □

Let

$$g_k \triangleq \sum_{i=k-\mathscr{T}+1}^{k} \left(y_i - C\hat{x}_{i|i-1}\right)^T \mathscr{P}^{-1} \left(y_i - C\hat{x}_{i|i-1}\right), \tag{2.11}$$

where $\mathscr{T}$ is the window size. Based on Theorem 1, it is known that when the system is operating normally, g_k has a χ^2 distribution with $m\mathscr{T}$ degrees of freedom[1], implying that there is lower probability that a larger g_k occurs. Therefore, the χ^2 detector at time k takes the following form:

$$g_k \underset{H_1}{\overset{H_0}{\lessgtr}} \eta, \tag{2.12}$$

where η is the threshold, usually chosen for a specific false alarm probability. If g_k is greater than the threshold, then the detector will trigger an alarm.

2.2.5 Attacker Model

It is assumed that a malicious third party wants to break the control system described above. The attacker is assumed to have the capability to perform the following actions:

1. He can inject an external control input u_k^a into the system.

2. Conservatively, he can read all sensor readings and modify them arbitrarily. The readings modified by the attacker are denoted by y'_k.

Given these capabilities, the attacker is assumed to implement an attack strategy, which can be divided into two stages:

1. The attacker records a sufficient number of y_ks without giving any input to the system.

2. The attacker gives a sequence of desired control input while replaying the previous recorded y_ks.

[1] The concept of degrees of freedom is a component of the definition of the χ^2 distribution. Please refer to Scharf and C. Demeure ([38]) for more details.

Remark 2. *It is important to note the lack on the part of the attacker to read the control inputs sent to the actuators. This assumption is vital in not disclosing the key in the cryptosystem — the watermarking signal. In the event that the attacker can read the control inputs, he might very well set up a fake system that takes these control inputs, generates the necessary measurements and sends them over to the controller, thereby completel dissociating the actual system and the controller. This is equivalent to a man-in-the-middle attack, and there is no way for the detector to know that such an attack has taken place. The only way to detect such an attack would be to introduce a "shared secret" between the controller and the system. However, the design of such a mechanism is out of the scope of the current work.*

The attack on the sensors can be executed by breaking the cryptography algorithm. Another way to perform an attack, which is thought to be much harder to defend, is to use physical attacks. For example, the readings of a temperature sensor can be manipulated if the attacker puts a heater near the sensor. Such kinds of attacks may be easy to carry out when sensors are spatially distributed in remote locations.

When the system is under attack, the controller cannot perform closed-loop control, since the sensory information is not available. Therefore, control performance of the system cannot be guaranteed during replay attack. The only way to counter such an attack is to detect it happening.

In the attacking stage, the goal of the attacker is to make the fake readings y_k's look like normal y_ks. Replaying the previous y_ks is just the easiest way to achieve this goal. There are other methods, such as machine learning or system identification, to generate a fake sequence of readings.

2.3 Previous Work

This section focuses on recapping the results previously accrued by Mo and Sinopoli ([9]).

2.3.1 Feasibility Of Attack

If we define $\mathscr{A} \overset{\Delta}{=} (A + BL)\,(I - KC)$, then it is proven by Mo and Sinopoli ([9]) that if $\mathscr{A}$ is stable, the distribution of g_k under replay attack will converge exponentially to the same distribution as g_k without the attack. As a result the asymptotic detection rate of the χ^2 detector is the same as its false alarm rate, i.e., the detector is unable to distinguish a system under the replay attack from a system that is running normally.

2.3.2 Countermeasure — Physical Watermarking

A watermark is a timeworn, well-established method of security and authentication, established in Italy during the thirteenth century. In its original sense, it is a recognizable image or text in a paper usually formed by thickness and/or density variations in the paper. The watermark can be discerned as a shaded pattern when viewed by reflected or transmitted light, which however, interferes only minimally or not at all with the printed or written matter on the paper. Watermarks are used to this day on banknotes, passports and postage stamps to prevent counterfeiting. Figure 2.1 shows a

Figure 2.1: A twenty euro banknote held against the light to show the watermark and the denomination. (Source: Wikimedia Foundation)

twenty-euro banknote held against the light to show the watermark and the denomination.

This principle of the physical watermark has been employed in recent years in the form of a digital signal to identify ownership and source of digital media like images, sound files and movies. Like traditional watermarks, digital watermarks are only discernible after applying some algorithm.

To detect a replay attack in the linear system under question, a small random authentication signal $\Delta u_k \sim \mathcal{N}(0, \mathcal{Q})$ is superimposed on the optimal control input u_k^*, which serves as a time stamp. It is proved that asymptotically the expectation of g_k under the attack will increase to

$$\lim_{k\to\infty} E\left[g_k\right] = m\mathscr{T} + 2\mathrm{tr}\left(C^T\mathscr{P}^{-1}C\mathscr{U}\right)\mathscr{T}. \tag{2.13}$$

where $\mathscr{U}$ is the solution of the Lyapunov equation

$$\mathscr{U} - B\mathscr{Q}B^T = \mathscr{A}\mathscr{U}\mathscr{A}^T. \tag{2.14}$$

The main problem of the combination of a Linear Quadratic Gaussian controller and a Kalman filter is that the whole control system is fairly static, which renders it vulnerable to a replay attack. In order to detect such a replay attack, one methodology is to redesign the control signal as

$$u_k = u_k^* + \Delta u_k, \tag{2.15}$$

where u_k^* is the optimal LQG control signal and the sequence Δu_k is drawn from an i.i.d. Gaussian distribution with zero mean and covariance $\mathscr{Q}$, and independent of u_k^*. Figure 2.2 shows the system diagram, including the attacker and the watermarking signal.

The sequence Δu_k acts as a time-stamped watermark, an authentication signal. It is chosen to be zero mean so as not to introduce any bias into the system. The presence of this extra authentication signal will cause the controller to not be optimal — in order to decrease the vulnerability of the system to the attack, the control performance must be sacrificed. Mo and Sinopoli

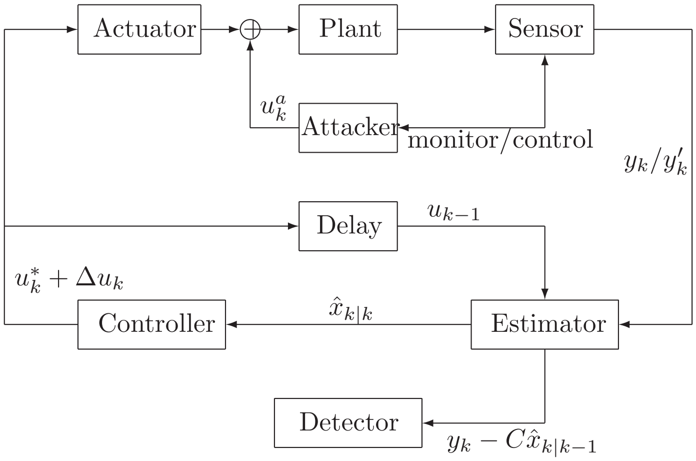


Figure 2.2: System Diagram with Physical Watermarking

([9]) proved that the increase in LQG cost (ΔJ) is $\operatorname{tr}\left(\left(U + B^T SB\right)\mathscr{Q}\right)$.

The remaining chapter section details the theoretical results beyond the work detailed above.

2.4 New Countermeasure — Using an Unstable $\mathscr{A}$

The feasibility result in [9] is that if $\mathscr{A}$ is unstable, then g_k goes to infinity exponentially fast, triggering the detector. One possible way to counter the replay attack is to redesign the control system, i.e. using non-optimal estimation and control gain matrices K and L, so that $\mathscr{A}$ becomes unstable while maintaining stability of the system. However, since K and L no longer remain optimal in the LQG sense, the LQG cost does increase.

The LQG cost for using non-optimal K and L can be characterized. It is known that

$$x_{k+1} = Ax_k + Bu_k + w_k = Ax_k + BL\hat{x}_k + w_k, \tag{2.16}$$

and

$$\begin{aligned}
\hat{x}_{k+1|k} &= A\hat{x}_k + Bu_k = (A + BL)\,\hat{x}_k \\
\hat{x}_{k+1} &= \hat{x}_{k+1|k} + K\left(y_{k+1} - C\hat{x}_{k+1|k}\right) \\
&= (I - KC)\,(A + BL)\,\hat{x}_k + Ky_{k+1} \\
&= (I - KC)\,(A + BL)\,\hat{x}_k + K\,(Cx_{k+1} + v_{k+1}) \\
&= KCAx_k + (A + BL - KCA)\,\hat{x}_k + KCw_k + Kv_{k+1}. \qquad (2.17)
\end{aligned}$$

Equation (2.16) and Equation (2.17) can be written in matrix form as

$$\begin{pmatrix} x_{k+1} \\ \hat{x}_{k+1} \end{pmatrix} = \begin{pmatrix} A & BL \\ KCA & A + BL - KCA \end{pmatrix} \begin{pmatrix} x_k \\ \hat{x}_k \end{pmatrix} + \begin{pmatrix} I \\ KC \end{pmatrix} w_k + \begin{pmatrix} 0 \\ K \end{pmatrix} v_{k+1}. \tag{2.18}$$

Let $\hat{A}$ be defined as

$$\hat{A} \triangleq \begin{pmatrix} A & BL \\ KCA & A + BL - KCA \end{pmatrix}. \tag{2.19}$$

Moreover, let $\hat{R}$ be defined as the covariance matrix of the second and third terms on the right hand side of Equation (2.18):

$$\hat{R} \triangleq \begin{pmatrix} I \\ KC \end{pmatrix} Q \begin{pmatrix} I & C^T K^T \end{pmatrix} + \begin{pmatrix} 0 \\ K \end{pmatrix} R \begin{pmatrix} 0 & K^T \end{pmatrix} \tag{2.20}$$

The LQG cost for non-optimal K and L can now be derived, which is given by the following theorem:

Theorem 3. *The LQG cost of using arbitrary estimation and control gain K and L is*

$$J = tr\left(\begin{pmatrix} W & 0 \\ 0 & L^T U L \end{pmatrix} \hat{Q} \right), \tag{2.21}$$

where $\hat{Q}$ is the solution of the following Lyapunov equation:

$$\hat{Q} = \hat{A}\hat{Q}\hat{A}^T + \hat{R}. \tag{2.22}$$

Proof. It is easy to see that since a fixed gain controller and estimator is used,

$$J = \lim_{k\to\infty} x_k^T W x_k + u_k^T U u_k, \tag{2.23}$$

which can be then written in matrix form as

$$\begin{aligned} J &= \lim_{k\to\infty} \begin{pmatrix} x_k^T & u_k^T \end{pmatrix} \begin{pmatrix} W & 0 \\ 0 & U \end{pmatrix} \begin{pmatrix} x_k \\ u_k \end{pmatrix} \\ &= \lim_{k\to\infty} \operatorname{tr}\left(\begin{pmatrix} W & 0 \\ 0 & U \end{pmatrix} \begin{pmatrix} x_k \\ u_k \end{pmatrix} \begin{pmatrix} x_k^T & u_k^T \end{pmatrix} \right) \\ &= \lim_{k\to\infty} \operatorname{tr}\left(\begin{pmatrix} W & 0 \\ 0 & L^TUL \end{pmatrix} \operatorname{Cov}\left(\begin{pmatrix} x_k \\ u_k \end{pmatrix} \right) \right). \end{aligned} \tag{2.24}$$

Let

$$\hat{Q} \triangleq \lim_{k\to\infty} \operatorname{Cov}\left(\begin{pmatrix} x_k \\ u_k \end{pmatrix} \right). \tag{2.25}$$

By Equation (2.18),

$$\operatorname{Cov}\left(\begin{pmatrix} x_{k+1} \\ u_{k+1} \end{pmatrix} \right) = \hat{A}\operatorname{Cov}\left(\begin{pmatrix} x_k \\ u_k \end{pmatrix} \right) \hat{A}^T + \hat{R}. \tag{2.26}$$

Taking the limit on both sides, $\hat{Q}$ becomes the solution of the following Lyapunov equation

$$\hat{Q} = \hat{A}\hat{Q}\hat{A}^T + \hat{R}.$$

Therefore, the LQG cost is given by

$$J = \operatorname{tr}\left(\begin{pmatrix} W & 0 \\ 0 & L^T U L \end{pmatrix} \hat{Q}\right).$$

□

There might not be enough freedom to redesign the control, which is required for this countermeasure to be implemented. However, the inclusion of this method is not just for the sake of completeness — as g_k increases exponentially, this method provides the highest asymptotic probability of detection, in the case that it *is* feasible.

2.5 Physical Watermarking

However, it is likely that the design constraints do not allow $\mathscr{A}$ to be unstable. This might be due to tight constraints on operating costs, safety parameters, etc. In such cases, the physical watermarking countermeasure can be applied. The results of Mo and Sinopoli [9] are extended, by providing a way to design the watermark for multi-input multi-output (MIMO) systems.

In a SISO system, there is only one way to insert the random signal, and only one way to observe it. Thus, to achieve a certain detection rate, a certain performance loss would have to be accepted. However, in the case of MIMO systems, the authentication signal can be inserted on one input or on many, with different strengths, independent or not.

The different possible forms of the signal can be better visualized using a vector interpretation of the different components — each control input can be considered as a coordinate in a p-dimensional space. The multivariate normal distribution that is characterized by the covariance matrix $\mathscr{Q}$ has equidensity contours that form ellipsoids in the p-dimensional space. The directions of the principal axes of the ellipsoids are given by the eigenvectors of $\mathscr{Q}$, and their squared relative lengths are given by the corresponding eigenvalues. It is possible that $\mathscr{Q}$ has less that p non-zero eigenvalues, in which case the ellipsoid would be infinitely thin in a particular direction. Figure 2.3 shows a possible ellipsoid for a system with 3 control inputs ($p = 3$).

The authentication signal Δu_k can be optimized such that the detection requirements are met while minimizing the effect on controller performance. Since the authentication signal has to be zero-mean, the design hinges on the covariance matrix $\mathscr{Q}$. Let the optimal value of $\mathscr{Q}$, based on the design requirements, be denoted by $\mathscr{Q}^*$.

The optimization problem can be set up in two ways. Firstly, the LQG performance loss (ΔJ) can be constrained to be less than some design parameter Θ, and the increase (Δg_k) in the expected value of the quadratic residues in case of an attack maximized. In this case, the optimal $\mathscr{Q}^*$ is the

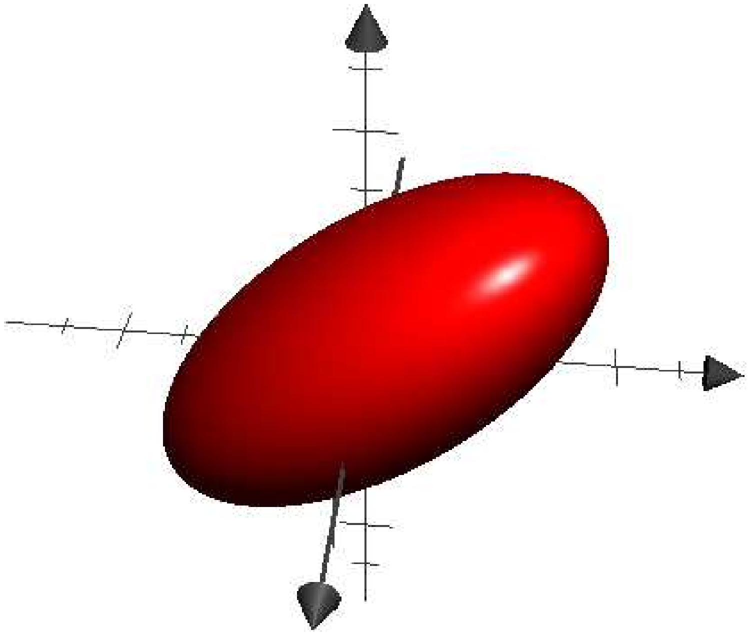


Figure 2.3: A geometric interpretation of the covariance matrix $\mathscr{Q}$ for $p = 3$. The principle axes of the ellipsoid are determined by the eigenvectors of $\mathscr{Q}$, and their relative lengths, by the corresponding eigenvalues.

solution to the optimization problem:

$$\begin{aligned} \underset{\mathscr{Q}}{\text{maximize}} \quad & \operatorname{tr}\left(C^T \mathscr{P}^{-1} C \mathscr{U}\right) \\ \text{subject to} \quad & \mathscr{U} - B\mathscr{Q}B^T = \mathscr{A}\mathscr{U}\mathscr{A}^T \\ & \mathscr{Q} \succeq 0 \\ & \operatorname{tr}\left[\left(U + B^T S B\right)\mathscr{Q}\right] \leq \Theta. \end{aligned} \tag{2.27}$$

Theorem 4. *There exists an optimal $\mathscr{Q}^*$ for Equation (2.27) of the following form:*

$$\mathscr{Q}^* = \alpha\omega\omega^T, \tag{2.28}$$

where $\alpha > 0$ is scalar and ω is a vector such that $\omega^T\omega = 1$.

Proof. Suppose that $\mathscr{Q}^*$ is the optimal solution of Equation (2.27) and $\mathscr{U}^*$ is the solution of

$$\mathscr{U}^* - B\mathscr{Q}^*B^T = \mathscr{A}\mathscr{U}^*\mathscr{A}^T. \tag{2.29}$$

Since $\mathscr{Q}^*$ is positive semidefinite, it is known that

$$\mathscr{Q}^* = \Omega \underbrace{\begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}}_{\Lambda} \Omega^T, \tag{2.30}$$

where $\lambda_i \geq 0$s are the eigenvalues of $\mathscr{Q}^*$ and $\Omega = (\omega_1, \omega_2, \ldots, \omega_p)$ is an orthonormal matrix, such that $\omega_i \in \mathbb{R}^p$. As a result, $\mathscr{Q}^*$ can be written as the sum of p rank 1 matrices:

$$\mathscr{Q}^* = \sum_{i=1}^{p} \lambda_i \omega_i \omega_i^T. \tag{2.31}$$

Let $\mathscr{Q}_i$ be defined as

$$\mathscr{Q}_i \triangleq \alpha_i \omega_i \omega_i^T, \tag{2.32}$$

where $\alpha_i > 0$ is chosen such that

$$\operatorname{tr}\left[\left(U + B^T S B\right) \mathscr{Q}_i\right] = \Theta. \tag{2.33}$$

Moreover, let $\mathscr{U}_i$ be defined as the solution of the following Lyapunov equation:

$$\mathscr{U}_i - B\mathscr{Q}_i B^T = \mathscr{A}\mathscr{U}_i\mathscr{A}^T. \tag{2.34}$$

It is clear that the optimal $\mathscr{Q}^*$ must satisfy

$$\operatorname{tr}\left[\left(U + B^T S B\right) \mathscr{Q}^*\right] = \Theta. \tag{2.35}$$

Therefore, since

$$\mathscr{Q}^* = \sum_{i=1}^{p} \frac{\lambda_i}{\alpha_i} \mathscr{Q}_i, \tag{2.36}$$

it can be seen that

$$\begin{aligned}\Theta &= \mathrm{tr}\left[\left(U + B^T S B\right)\mathscr{Q}^*\right] \\ &= \sum_{i=1}^{p} \frac{\lambda_i}{\alpha_i} \mathrm{tr}\left[\left(U + B^T S B\right)\mathscr{Q}_i\right] \\ &= \sum_{i=1}^{p} \frac{\lambda_i}{\alpha_i} \Theta, \end{aligned} \tag{2.37}$$

which proves that

$$\sum_{i=1}^{p} \frac{\lambda_i}{\alpha_i} = 1. \tag{2.38}$$

Furthermore, it is easy to see that since Lyapunov equation is linear,

$$\mathscr{U}^* = \sum_{i=1}^{p} \frac{\lambda_i}{\alpha_i} \mathscr{U}_i. \tag{2.39}$$

Hence,

$$\mathrm{tr}\left(C^T \mathscr{P}^{-1} C \mathscr{U}^*\right) = \sum_{i=1}^{p} \frac{\lambda_i}{\alpha_i} \mathrm{tr}\left(C^T \mathscr{P}^{-1} C \mathscr{U}_i\right). \tag{2.40}$$

As a result, $\mathscr{Q}^*$ is a convex combination of p feasible $\mathscr{Q}_i$s. Since $\mathscr{Q}^*$ is optimal, we know that for any $\lambda_i > 0$, the corresponding $\mathscr{Q}_i$ must also be optimal, which finishes the proof. □

Going back to the geometric visualization, this theorem states that the ellipsoid associated with $\mathscr{Q}$ will always have only one non-zero principal axis. In essence, instead of an ellipse, the optimal $\mathscr{Q}$ can be denoted by a p dimensional vector, the direction of which is characterized by the form of $\mathscr{Q}$, and the length of which is dependent on the norm of $\mathscr{Q}$.

The fact that $\mathscr{Q}^*$ has rank 1 has a direct bearing on the computation requirement. The number of independent random noise generators required is equal to the rank of $\mathscr{Q}^*$. Naïvely, one would have to use one independent random noise generator per system input, in order to protect all of them. However, irrespective of the number of system inputs, the rank of $\mathscr{Q}^*$ is always 1, which means that a single random noise generator will suffice for a system with any number of inputs. This also implies that only one random noise generator needs to be included in the "trusted base" of the controller hardware and software.

Ideally, if there were a design constraint on the LQG cost, one would try to optimize the detection rate. However, it can be shown that under attack g_k follows a generalized χ^2 distribution, and no analytical form for the detection rate can be accrued. Thus, only the maximization of the expectation in the case of an attack is attempted, with the intuition that the detection rate in such a case will be close to the maximum possible.

It can be seen from results in [9] that the increase (ΔJ) in LQG cost and increase (Δg_k) in the expectation of the quadratic residues are linear functions of the noise covariance matrix $\mathscr{Q}$. Thus the optimization problem is a semi-definite programming problem and hence can be solved efficiently. Furthermore, it can be seen that if the constraints are changed from Θ to $\alpha\Theta$, the optimal $\mathscr{Q}^*$ will be changed to $\alpha\mathscr{Q}$.

Another way of optimizing is to constrain the increase (Δg_k) in the expected values of the quadratic residues to be above a fixed value Γ, thereby guaranteeing a certain rate of detection, and the performance loss (ΔJ) can be minimized. The optimal $\mathscr{Q}^*$ is now the solution to the optimization problem:

$$\begin{aligned} \underset{\mathscr{Q}}{\text{minimize}} \quad & \operatorname{tr}\left[\left(U + B^T S B\right) \mathscr{Q}\right] \\ \text{subject to} \quad & \mathscr{U} - B\mathscr{Q}B^T = \mathscr{A}\mathscr{U}\mathscr{A}^T \\ & \mathscr{Q} \succeq 0 \\ & \operatorname{tr}\left(C^T \mathscr{P}^{-1} C \mathscr{U}\right) \geq \Gamma. \end{aligned} \tag{2.41}$$

2.6 Decoupling the Design Problem

The solutions of the two optimization problems given in Equations 2.27 and 2.41 will be scalar multiples of each other, thus solving either optimization problem guarantees same performance. An intuitive way to see this is that $\mathscr{Q}^*$ measures the sensitivity of the system output to the different inputs, thus making it a system property.

These properties can be applied to decouple the design of the signal into two steps, *Form* and *Norm*.

1. *Form of $\mathscr{Q}$* — The structure of the matrix $\mathscr{Q}$ is a system property, and can be ascertained for any value of the thresholds (Θ or Γ).

2. *Norm of $\mathscr{Q}$* — The norm of $\mathscr{Q}$ can be designed in the second step, taking into performance the required detector performance (by using a linear

multiplier to limit the quadratic residues to be above the threshold Γ), or the required control performance (by using a linear multiplier to limit the LQG performance loss to be less than the threshold Θ)

The first step of this approach requires setting up and solving an optimization problem, which, although technically is a linear programming problem, can be significantly large, involving as it does matrices of size $n \times n$, where n is the number of internal states of a system. The number of internal states of a system can indeed be quite large, especially if the system involves some kind of a physical delay, which is usually the case. A physical delay creates a system that is no longer memory-less, and the number of memory states required in the system is of the order of the number of discrete time steps that make up the maximum temporal delay in the system. For example, in the simulations of chapter 3, the system involves a delay of 6 minutes, which, using a discretization time-step of 0.01 minutes, translates to around 600 memory states. The optimization problem, thus, involves matrices of size more than 600×600, which can take a significant time to solve.

However, once the optimization is set up and solved offline, a receiver operating characteristic curve can be generated for possible norms of $\mathscr{Q}$. As shown in Figure 2.4, as $\|\mathscr{Q}\|$ increases, the ROC curve tends towards the optimal point $(\alpha = 0, \beta = 1)$. An operating point for the detector can be chosen by first choosing strength of the signal, and then a detection threshold.

In case of an increased threat level, the security of the system can be increased by "turning a single dial", i.e., changing the variance of the single random noise generator in the system.

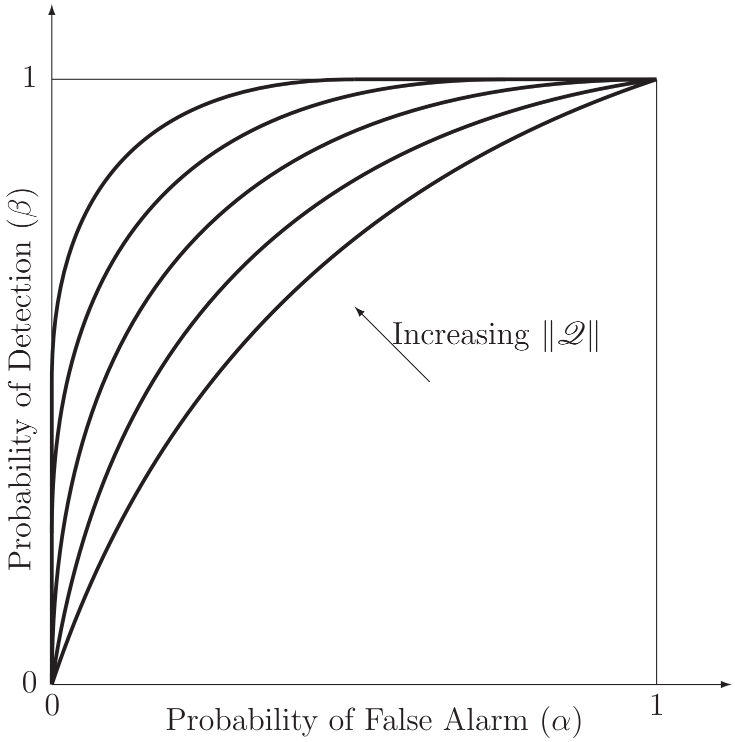


Figure 2.4: Possible Receiver Operating Characteristics for different norms of covariance matrix $\mathscr{Q}$

2.7 New Countermeasure — Cross-correlator Detector

The Kalman filter, LQG controller and the chi^2-detector all utilize the zero-mean Gaussian nature of the process and measurement noise. The chief reasons for using these three is their inter-compatibility, and the ease of theoretical analysis. The key idea behind physical watermakring, however, can be applied irrespective of the choice of estimator, controller, and detector

— the nature of dependence of the detection rate and strength of noise added will remain the same, even though the actual expressions will change.

As an example, in this section we will consider a detector that takes the cross-correlation of the expected measurements and the actual measurements accrued from the sensors.

Implementing the χ^2 detector requires the implementation of a Kalman estimator. However, in some systems, a Kalman estimator might not be feasible, due to noise characteristics or system observability. The noisy-control countermeasure, however, can still be applied, to virtually any controller and any detector, as long as a virtual system can be implemented. A signal $\Delta u_k \sim \mathcal{N}(0, \sigma^2)$ is added to the control signal. The effect of the control input on the virtual system can be calculated, and the outputs compared.

Although the implementation is applicable for any estimator, controller, and detector, for comparative purposes, the Kalman-LQG system from the previous subsection used, with the cross-correlator detector to derive the characteristics of this countermeasure. The system evolution equation is:

$$\begin{pmatrix} x_{k+1} \\ \hat{x}_{k+1} \end{pmatrix} = \underbrace{\begin{pmatrix} A & BL \\ KCA & A + BL - KCA \end{pmatrix}}_{\hat{A}} \begin{pmatrix} x_k \\ \hat{x}_k \end{pmatrix} + \underbrace{\begin{pmatrix} B \\ B \end{pmatrix}}_{\hat{B}} \Delta u_k + \begin{pmatrix} I \\ KC \end{pmatrix} w_k + \begin{pmatrix} 0 \\ K \end{pmatrix} v_{k+1}, \tag{2.42}$$

and the measurement equation is:

$$y_k = \underbrace{\begin{pmatrix} C & 0 \end{pmatrix}}_{\hat{C}} \begin{pmatrix} x_k \\ \hat{x}_k \end{pmatrix} + v_k. \tag{2.43}$$

Note that $\hat{A}$ is the same as defined in Equation (2.19). For the virtual system, the system evolution equation is:

$$\begin{pmatrix} x'_{k+1} \\ \hat{x}'_{k+1} \end{pmatrix} = \hat{A} \begin{pmatrix} x'_k \\ \hat{x}'_k \end{pmatrix} + \hat{B}\Delta u'_k + \begin{pmatrix} I \\ KC \end{pmatrix} w'_k + \begin{pmatrix} 0 \\ K \end{pmatrix} v'_{k+1}, \tag{2.44}$$

and the measurement equation is:

$$y'_k = \hat{C} \begin{pmatrix} x'_k \\ \hat{x}'_k \end{pmatrix} + v'_k. \tag{2.45}$$

It is assumed that $x_0 \sim \mathcal{N}(\bar{x}_0, \Sigma)$, $x'_0 \sim \mathcal{N}(\bar{x}_0, \Sigma)$, $\Delta u \sim \mathcal{N}(0, \mathscr{Q})$, $w_k \sim \mathcal{N}(0, Q)$, $w'_k \sim \mathcal{N}(0, Q)$, $v_k \sim \mathcal{N}(0, R)$, and $v'_k \sim \mathcal{N}(0, R)$ are all independent of each other. Let the detector run another virtual system, which is connected directly to the controller and cannot be attacked by the attacker.

$$\begin{pmatrix} x''_{k+1} \\ \hat{x}''_{k+1} \end{pmatrix} = \hat{A} \begin{pmatrix} x''_k \\ \hat{x}''_k \end{pmatrix} + \hat{B}\Delta u_k + \begin{pmatrix} I \\ KC \end{pmatrix} w''_k + \begin{pmatrix} 0 \\ K \end{pmatrix} v''_{k+1}, \tag{2.46}$$

and the measurement equation is:

$$y''_k = \hat{C} \begin{pmatrix} x''_k \\ \\ \hat{x}''_k \end{pmatrix} + v''_k. \tag{2.47}$$

Consider the detector variable $g_k = y'^T y'' = \mathrm{tr}\left(y' y''^T\right)$. It can be proved that, in the absence of a replay attack,

$$E\left[y' y''^T\right] = \hat{C} \mathscr{R} \hat{C}^T, \tag{2.48}$$

where $\mathscr{R}$ is the solution of the following Lyapunov equation:

$$\hat{A} \mathscr{R} \hat{A}^T + \hat{B} \mathscr{Q} \hat{B}^T = \mathscr{R}. \tag{2.49}$$

If the attacker replays the outputs y, or if he is running another virtual system, the $\Delta u'$ generated by the attacker will be independent of the Δu used in the controller's virtual system. In case of either form of attack, $\mathscr{R}$ becomes 0, causing $E\left[y' y''^T\right]$ to drop to 0 as well. We can thus detect the absence of the authentication signal in the output and hence, the attack.

Similar to the χ^2 detector, in the case of MIMO systems, the covariance matrix $\mathscr{Q}$ can be optimized, such that the detection requirements are met while minimizing the effect on controller performance. Just like the previous case, the optimization problem can be set up in two ways. Firstly, the LQG performance loss (ΔJ) can be constrained to be less than some design parameter Θ, and the increase (Δg_k) in the expected value of the correlator output in case of an attack maximized. In this case, the optimal $\mathscr{Q}^*$ is the solution to the optimization problem:

$$\begin{aligned} \underset{\mathscr{Q}}{\text{maximize}} \quad & \operatorname{tr}\left(\hat{C}\mathscr{R}\hat{C}^T\right) \\ \text{subject to} \quad & \hat{A}\mathscr{R}\hat{A}^T + \hat{B}\mathscr{Q}\hat{B}^T = \mathscr{R} \\ & \mathscr{Q} \succeq 0 \\ & \operatorname{tr}\left[\left(U + B^TSB\right)\mathscr{Q}\right] \leq \Theta. \end{aligned} \tag{2.50}$$

Secondly, the increase (Δg_k) in the expected values of the quadratic residues can be constrained to be above a fixed value Γ, thereby guaranteeing a certain rate of detection, and the performance loss (ΔJ) can be minimized. The optimal $\mathscr{Q}^*$ is now the solution to the optimization problem:

$$\begin{aligned} \underset{\mathscr{Q}}{\text{minimize}} \quad & \operatorname{tr}\left[\left(U + B^TSB\right)\mathscr{Q}\right] \\ \text{subject to} \quad & \hat{A}\mathscr{R}\hat{A}^T + \hat{B}\mathscr{Q}\hat{B}^T = \mathscr{R} \\ & \mathscr{Q} \succeq 0 \\ & \operatorname{tr}\left(\hat{C}\mathscr{R}\hat{C}^T\right) \geq \Gamma. \end{aligned} \tag{2.51}$$

Theorem 5. *There exists and optimal $\mathcal{Q}^*$ for Equation* (2.50) *of the following form:*

$$\mathcal{Q}^* = \alpha\omega\omega^T, \tag{2.52}$$

where $\alpha > 0$ is a scalar and ω is a vector with $\omega^T\omega = 1$.

Proof. The proof is very similar to that of Theorem 4, hence is omitted. □

Remark 6. *Like the χ^2 detector, only the maximization of the expectation is attempted. The optimization problems are linear, and generate optimal $\mathscr{Q}^*$s which are multiples of each other.*

Chapter 3

Replay Attack: Example

In this section, a system that requires defense against the proposed replay attack is introduced. The countermeasures discussed in Section 2 are successively applied to the system. The importance of optimizing the signal is indicated by highlighting the differences in using unoptimized and optimized authentication signals.

3.1 Problem Formulation — Chemical Plant

The above methodology is applied to a simplified version of the Tennessee Eastman Control Challenge Problem ([39]). The original problem requires coordination of three unit operations, with 41 measured output variables (with added measurement noise) and 12 manipulated variables. The control challenge presented by this case study is quite complex. However, a simplified version was proposed by N. Lawrence Ricker in 1993 ([40]), which is the model we adopt. In this paper, Ricker derives a linear time-invariant dynamic model of the plant in its base-state, and a corresponding robust controller, with four outputs and four inputs[1]:

$$\mathbf{y} = \begin{pmatrix} F_4 \\ P \\ y_{A3} \\ V_L \end{pmatrix} = \mathbf{G}\mathbf{u} = \begin{pmatrix} g_{11} & 0 & 0 & g_{14} \\ g_{21} & 0 & g_{23} & 0 \\ 0 & g_{32} & 0 & 0 \\ 0 & 0 & 0 & g_{44} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix}. \tag{3.1}$$

The individual transfer functions are given in Equations 3.2–3.7 (the unit of s is assumed to be hr^{-1}):

$$g_{11} = \frac{1.7}{0.75s + 1}, \tag{3.2}$$

$$g_{21} = \frac{45\,(5.667s + 1)}{2.5s^2 + 10.25s + 1}, \tag{3.3}$$

$$g_{23} = \frac{-15s - 11.25}{2.5s^2 + 10.25s + 1}, \tag{3.4}$$

$$g_{32} = \frac{1.5}{10s + 1} e^{-0.1s}, \tag{3.5}$$

$$g_{14} = \frac{-3.4s}{0.1s^2 + 1.1s + 1}, \tag{3.6}$$

$$g_{44} = \frac{1}{s + 1}. \tag{3.7}$$

The system is sampled at 100 samples per minute. The values of Q, R, W, and U used for the controller are $Q = 0.01I$, $R, W, U = I$.

[1] The transfer function g_{23} is not given in [40]. It was estimated using the method described in the paper.

3.2 Attack Methodology

The attacker is considered to know the readings of all the sensors, with the ability to hijack and modify them, but not the dynamics of the system. The requirement of control over all sensors can be weakened if the system can be decomposed into several weakly coupled subsystems, compromising sensors for one subsystem may be sufficient. The only known fact is that the system is expected to be in steady state for the duration of the attack. Of the 30 minutes for which the system is simulated, the attacker records the sensor readings for the first fifteen minutes, and replays them to the controller for the next fifteen. The attack consists for varying the control inputs of the plant, to try and evolve it into a potentially dangerous state. Since no information from the system is conveyed to the controller, the system becomes open loop, without guarantees on control performance. The only way to get the system back into the controlled state is to detect and mitigate the attack.

3.3 Results

The system is initially simulated without any countermeasure to prove the feasibility of attack. In the next set of simulations, the physical watermarking countermeasure is introduced, in both the optimal and non-optimal forms. The methodology for designing the optimal *form* and *norm* is illustrated. Finally, the cross-correlator countermeasure is applied.

3.3.1 Feasibility of Attack

For the chemical plant, a W and U were chosen such that $\mathscr{A}$ is stable. A χ^2 detector with a window size of 10 samples (1 minute) is used. Figure 3.1a shows the value of g_k for a χ^2 detector, for the duration of 30 minutes, when no attack is present. Figure 3.1b shows the value of g_k when an attack occurs after the first 15 minutes. It can be seen that there is no appreciable statistical difference in g_k when an attack is present, making detection impossible.

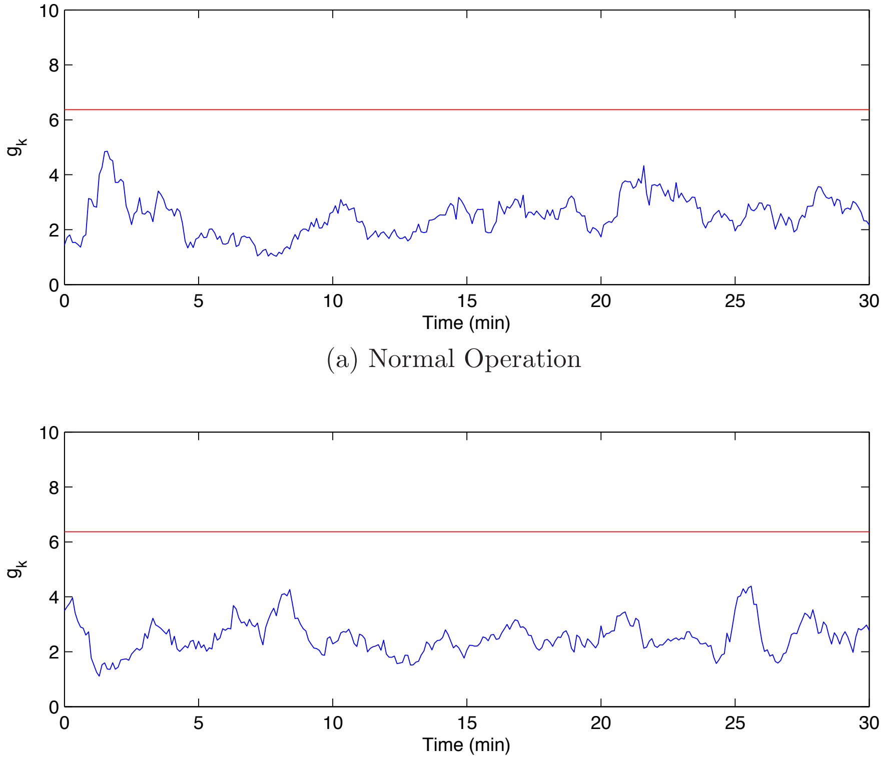
Thus, executing the attack without being detected is feasible.

3.3.2 Unstable $\mathscr{A}$

It is assumed that the design parameters are flexible enough to allow $\mathscr{A}$ to be unstable. K and L are generated randomly such that they form a good estimator-controller pair, such that $\mathscr{A}$ is unstable. A χ^2 detector with a window size of 10 samples (1 minute) is used. Figure 3.2 shows the value of g_k in normal operation and when an attack occurs after the first 15 minutes. It can be seen that the instability in $\mathscr{A}$ causes a change in g_k when an attack is present, which can be detected.

3.3.3 χ^2 Detector, Non-Optimal

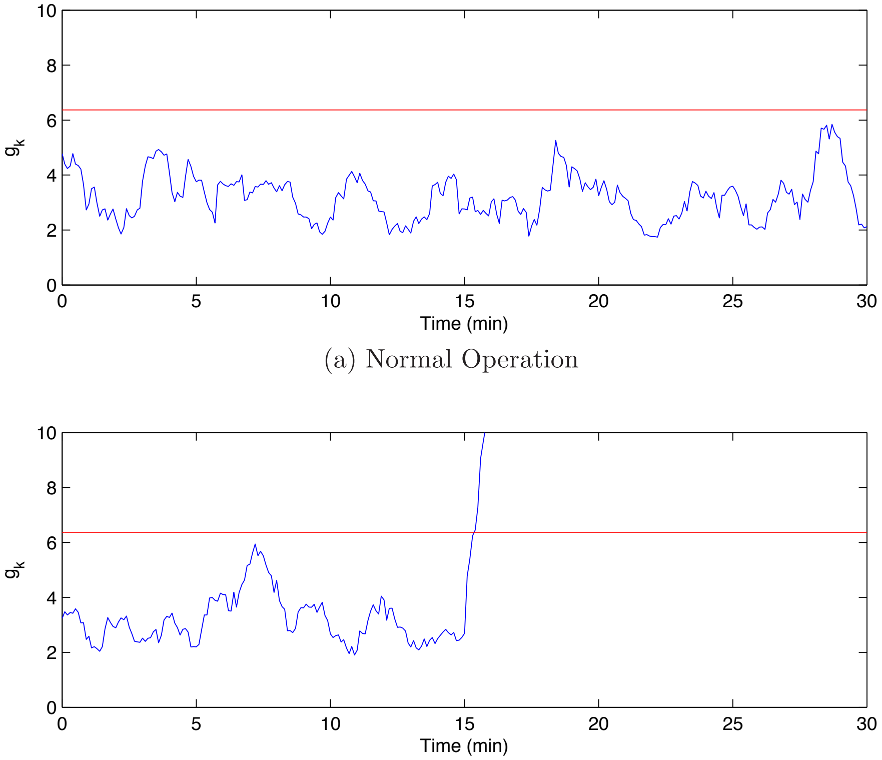
For this simulation, the estimator and controller are reverted to the original case of section 3.3.1. The countermeasure of “noisy-control” is now used for the system. A χ^2 detector with a window size of 10 samples (1 minute) is implemented. In this case, the authentication signal is not optimized. The



(a) Normal Operation

(b) Replay Attack

Figure 3.1: g_k as a function of time during normal operation, and a replay attack. This shows that the detector (with threshold at 99% shown) fails to detect the fall in g_k due to an attack.

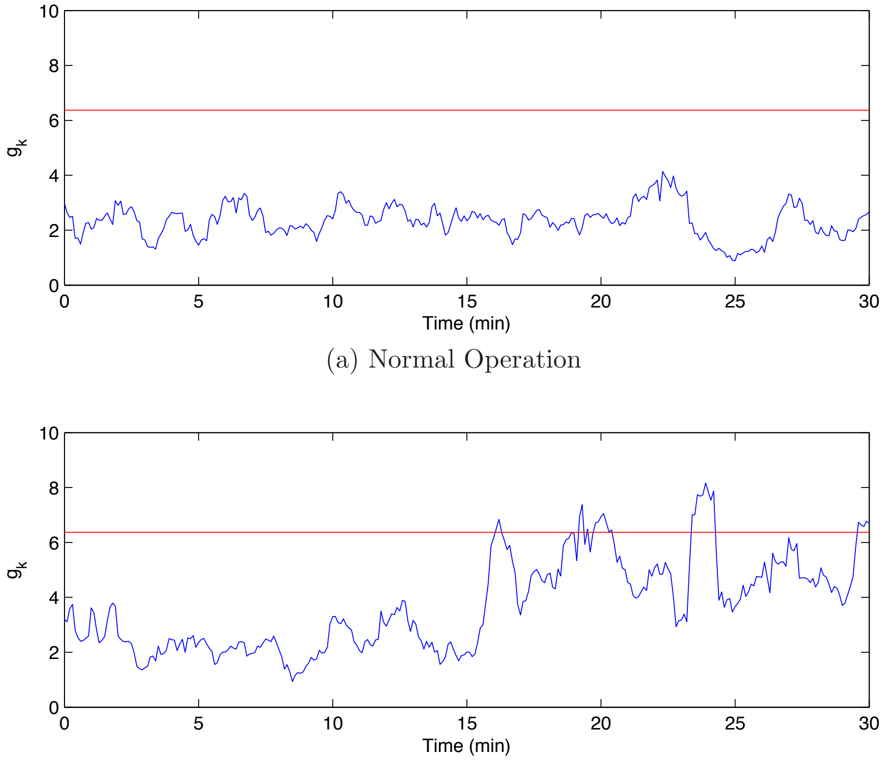


(a) Normal Operation

(b) Replay Attack

Figure 3.2: g_k as a function of time during normal operation, and a replay attack, using a controller with unstable $\mathscr{A}$. This shows that the detector (with threshold at 99% shown) is able to detect the fall in g_k due to an attack.

expected increase in LQG cost is 10% of the optimal LQG cost. In this case Figure 3.3a shows the value of g_k for a χ^2 detector, for the duration of 30 minutes, when no attack is present. Figure 3.3b shows the value of g_k when an attack occurs after the first 15 minutes. It can be seen that there is some difference in the statistical distribution of g_k with and without an attack.



(a) Normal Operation

(b) Replay Attack

Figure 3.3: g_k as a function of time during normal operation, and a replay attack. This shows that the detector (with threshold at 99% shown) is able to detect the fall in g_k due to an attack.

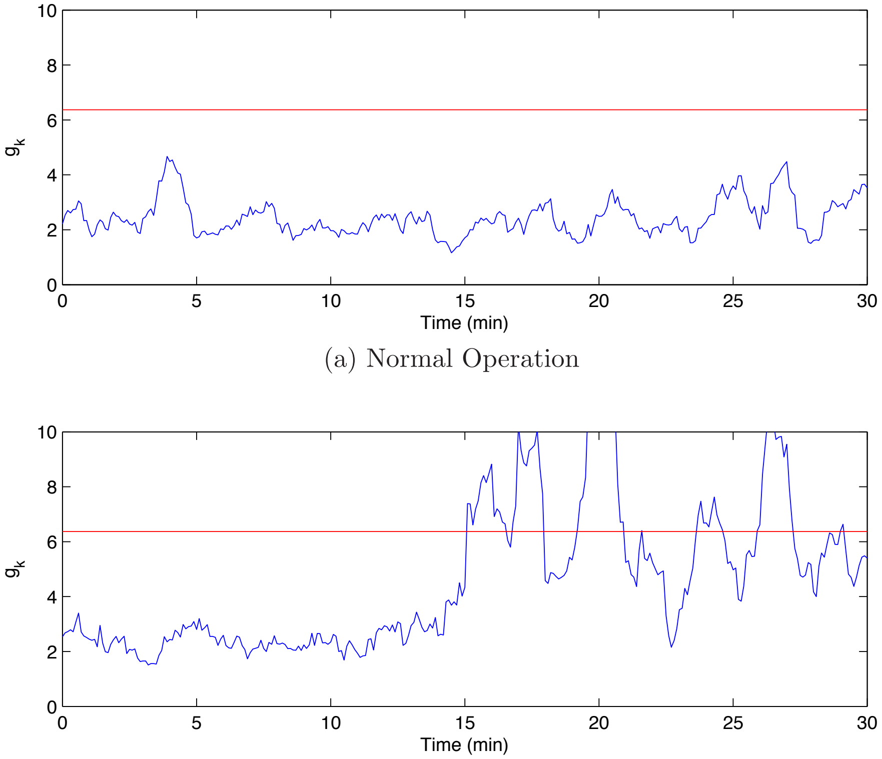
3.3.4 χ^2 Detector, Optimal

This simulation is similar to the one in section 3.3.3, except that the authentication signal is optimized such that the expected increase in LQG cost is 10% of the optimal LQG cost. In this case Figure 3.4a shows the value of g_k for a χ^2 detector, for the duration of 30 minutes, when no attack is present. Figure 3.4b shows the value of g_k when an attack occurs after the first 15 minutes. It can be seen there is significant difference in the statistical distribution of g_k with and without an attack. The results of this simulation, when compared to those of section 3.3.3, show the importance of optimizing the form of $\mathscr{Q}$.

In the next set of simulations, $\mathscr{Q}$ is scaled by 0.2, 0.4, 0.6, 0.8 and 1, which corresponds to setting Θ to 2%, 4%, 6%, 8%, and 10% respectively. A sample set of 500 simulations was carried out to calculate the Receiver Operating Characteristic (ROC) curve for each signal strength. These curves are shown in Figure 3.5. In this case, probability of detection 1 minute after the onset of the attack has been considered. It is easy to see that the performance of the detector improves with increase in $\|\mathscr{Q}^*\|$, so an appropriate signal strength can be designed considering the trade-off between the required ROC curve and allowed performance loss.

3.3.5 Cross-Correlator Detector, Optimal

In this simulation, we use a cross-correlator detector with a window size of 30 samples (3 minutes) and the authentication signal is optimized such that



(a) Normal Operation

(b) Replay Attack

Figure 3.4: g_k as a function of time during normal operation, and a replay attack. This shows that the detector (with threshold at 99% shown) is able to detect the fall in g_k due to an attack. Compared to Figure 3.3b, the change in the statistics of the signal upon attack is much more significant, and is less likely to be attributed to parameter change or inaccurate system knowledge.

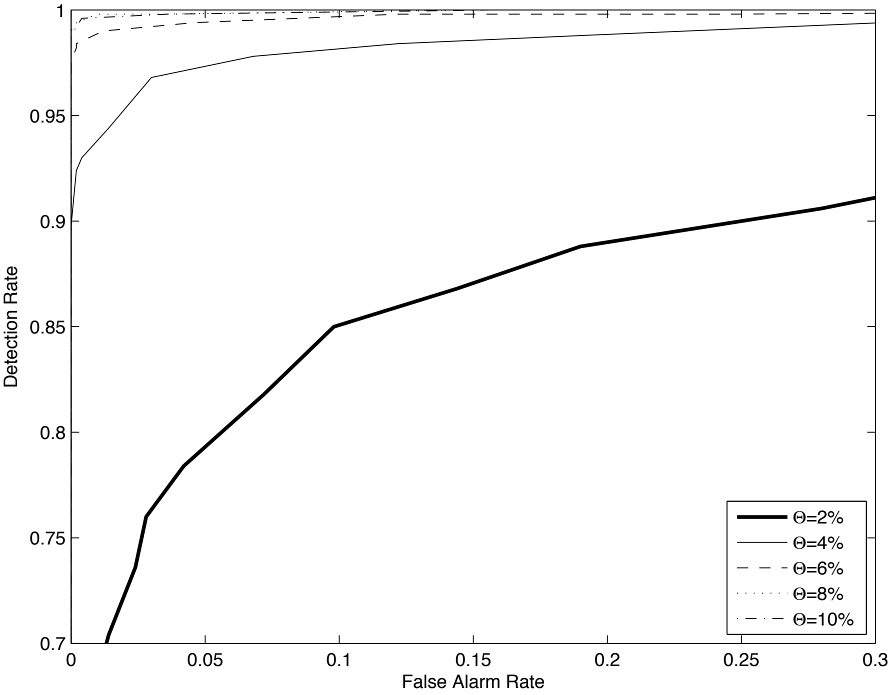
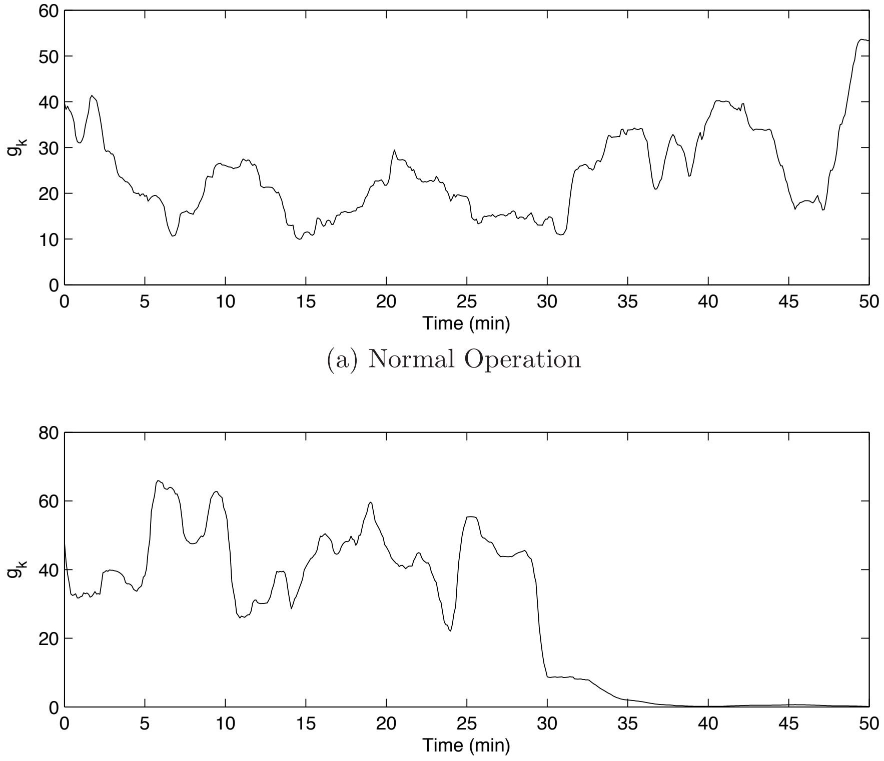


Figure 3.5: ROC curves for detector, when Θ is 2% (dark solid line), 4% (thin solid line), 6% (dashed line), 8% (dotted line) and 10% (dash-dot line). Detection up to 1 second after attack is considered.

the expected increase in LQG cost is 20% of the optimal LQG cost. The expected value the correlator output g_k is 30.996. Figure 3.6a shows the correlator output, for the duration of 30 minutes, when no attack is present. Figure 3.6b shows the correlator output when an attack occurs after the first 15 minutes. It can be seen that g_k drops significantly when an attack is in progress.



(a) Normal Operation

(b) Replay Attack

Figure 3.6: g_k as a function of time during normal operation, and a replay attack. This shows that the detector is able to detect the fall in g_k due to an attack.

Chapter 4

Integrity Attacks: Example

This section introduces the problem of economic dispatch in electrical power grids. The attack methodology is updated using state-of-the-art sensors as well as state-of-the-art attacking tools. The extent of disruption such an attack can cause while circumventing the current security measures is indicated.

4.1 Phasor Measurement Units

Phasor Measurement Units (PMUs) are devices that measure the various synchrophasors at each bus. Synchrophasors are voltage and current phasors measured synchronously at widely dispersed locations on power grid, which can be compared in real-time. These synchrophasors improve upon traditional state estimation calculated using unsynchronized data points collected every 2–4 seconds. Dubbed as “the MRI of our Power System”[1], Phasor Measurement Units, designed to measure these synchrophasors, were

[1] Power Grid Corporation of India, Limited

invented in 1988 at Virginia Polytechnic Institute and State University, by Dr. Arun G. Phadke and Dr. James S. Thorp. PMUs deliver 10–30 synchronous reports per second, and the necessary ±500 ns accuracy is provided by GPS time stamping.

PMUs are protected against loss of GPS signal, unintentional or otherwise, by the use internal reference clock for several seconds. However, GPS broadcasts can be spoofed without jamming. Practicality of GPS spoofing was established by the work of Prof. Brumley et al, Carnegie Mellon University among others. Such an attack involves fabricating a counterfeit signal from a GPS satellite, and placing an antenna to ensure fake signal drowns out real one. A properly orchestrated attack on a PMU will change time-stamps on PMU measurements, and hence the phase measurements.

4.2 Electricity Markets

In a wholesale electricity market such as found in many countries in the world, competing generators bid on supplying electricity to retailers, who then re-price it and sell it to consumers. For a wholesale electricity market to be economically efficient, it needs a coordinated spot market that carries out a "bid-based, security constrained, economic dispatch with nodal prices".

The day-ahead market determines the system prices by equating supply and demand — matching bids from generators and consumers at each node. The theoretical price at each node is the marginal cost of an additional unit of electricity to the system resulting from optimized redispatch of available electricity. This “shadow price” of the hypothetical kilowatt-hour of electricity is known as the locational marginal pricing and is used in some deregulated markets including the Pennsylvania-New Jersey-Maryland Interconnection[2] and New Zealand.

When network constraints, such as line limits being reached or exceeded, or contingencies such as generator failure or transformer outage occurring, costlier generation needs to be dispatched on the downstream side of the congestion, causing the nodal prices on either end of the constraint to diverge. The violation of constraints can only be determined by state estimation using measurements from the SCADA system. Thus, the fidelity of the ex-post settlement price for all market participants is based on the integrity of the state estimation.

Xie et al ([10]) studied the economic impact of a potential class of integrity cyber attacks on electric power market operations. They showed that with the knowledge of the transmission system topology, attackers might circumvent the bad data detection algorithms equipped in today’s state estimator. This, in turn, may be leveraged by attackers for consistent financial arbitrage such as virtual bidding at selected pairs of nodes.

4.3 Problem Formulation — Economic Dispatch

The notations used for the problem formulation are summarized in Table 4.1.

[2] serving all or parts of Delaware, Illinois, Indiana, Kentucky, Maryland, Michigan, New Jersey, North Carolina, Ohio, Pennsylvania, Tennessee, Virginia, West Virginia and the District of Columbia

i	Index for generator i
j	Index for load bus j
l	Index for transmission line l
k	Time k
I	Total number of generators
J	Total number of load buses
L	Total number of transmission lines
Ld_j	Load at bus j during run time
Pg_i	Generation at i during run time
x	A vector consisting of all Pg_i and Ld_j
z	Collection of sensor measurements
$C_i\left(Pg_i\right)$	Generation cost for producing Pg_i
$Pg_i^{\min\ (\max)}$	Minimum (maximum) available power from generator i
λ_i	Electricity price at bus i
F_l	Transmission flow at line l
$F_l^{\max}$	Maximum allowed transmission at line l
$F_l^{\min}$	Minimum allowed transmission at line l

Table 4.1: Notations used for problem formulation ([10])

The power market operates in three phases:

1. *Ex-Ante:* The ex-ante real-time market, which usually takes place every 10 to 15 minutes prior to real time, conducts security-constrained economic dispatch to determine the optimal power generation given the expected load:

$$\begin{aligned} \underset{Pg_i^*}{\text{maximize}} \quad & \sum_{i=1}^{I} C_i\left(Pg_i^*\right) && (4.1) \\ \text{subject to} \quad & \sum_{i=1}^{I} Pg_i^* = \sum_{j=1}^{J} Ld_j^* \\ & Pg_i^{\min} \leq Pg_i^* \leq Pg_i^{\max} \ \forall i = 1, 2, \dots, I \\ & F_l^{\min} \leq F_l^* \leq F_l^{\max} \ \forall l = 1, 2, \dots, L. && (4.2) \end{aligned}$$

   Based on the linearized DC power flow model, the line flow vector is a linear function of the nodal injection vector

$$F = H \begin{pmatrix} Ld \\ Pg \end{pmatrix}. \qquad (4.3)$$

2. *State Estimation:* Due to the stochastic nature of the demand Ld_j, the real-time values of Pg, Ld, and F may differ from the optimal values calculated in the ex-ante market clearing. Hence, measurements are necessary to estimate the real-time state variables. The real-time

system states (x) differ from the steady state values x^* :

$$x = x^* + w,\ F = H\left(x^* + w\right), \tag{4.4}$$

where w is a Gaussian random variable with zero mean and covariance Q. Since the SCADA system measures the nodal injection vectors as well as the line flows, the observation equation is:

$$z = \underbrace{\begin{pmatrix} I \\ H \end{pmatrix}}_{C} x + e, \tag{4.5}$$

where e is the measurement error, also assumed to be Gaussian with zero mean and covariance R. Since the observation equations and flow model are assumed to be linear, the solution of the minimum mean square error estimator is given by

$$\hat{x} = \underbrace{\left(C^T R^{-1} C\right)^{-1} C^T R^{-1}}_{P} z \tag{4.6}$$

3. *Bad Data Detection:* The bad data detection system implemented in state estimators compares the accrued measurements (z) with the expected measurements of a physical model. The residue r is defined as

$$r \triangleq z - C\hat{x}. \tag{4.7}$$

The detector triggers an alarm based by comparing the norm of r with certain threshold.

4. *Ex-Post:* Since the run time state variables Pg, Ld, and F are different from the dispatch level in ex-ante market, RTOs will calculate the vector of LMPs based on the run-time data for settlement purposes. The ex-post pricing model is described in detail by Li et al ([41]). If the positive and negative congestion sets are defined as:

$$cl_+ = \left\{ l \middle| \hat{F}_l \geq F_l^{\max} \right\},\ cl_- = \left\{ l \middle| \hat{F}_l \leq F_l^{\min} \right\}, \tag{4.8}$$

the ex-post formulation solves the SCED to obtain the LMPs for settlement:

$$\begin{aligned} \underset{Pg_i^*}{\text{maximize}} \quad & \sum_{i=1}^{I} C_i \left(\Delta Pg_i + \hat{Pg}_i^* \right) & (4.9) \\ \text{subject to} \quad & \sum_{i=1}^{I} \Delta Pg_i = 0 \\ & \Delta Pg_i^{\min} \leq \Delta Pg_i \leq \Delta Pg_i^{\max}\ \forall i = 1, 2, \ldots, I \\ & \Delta F_l \leq 0\ \forall \in cl_+ \\ & \Delta F_l \geq 0\ \forall \in cl_-. & (4.10) \end{aligned}$$

After solving the above optimization problem and computing the Lagrangian multipliers λ, η_l ζ_l, the nodal price at each load bus of the network, is defined as

$$\lambda_j = \lambda + H_j^T \left(\eta - \zeta\right), \tag{4.11}$$

where H_j is the jth column of the H matrix.

4.4 Timing Attacks

From an attacker's point of view, a Phasor Measurement Unit has several possible attack vectors — a network attack on the communication to the data concentrator, an attack that injects current locally to distort the phasor measurement, or an attack on the GPS unit. A current injection attack can be considered to be beyond the realm of possible attacks, since the current source needed to distort measurement would be too massive to utilize discretely. A network attack can be prevented by using sufficiently good encryption, and is out of the scope of this work.

A timing attack that breaks the synchronicity of the phasor measurements could be a major problem for PMUs. As per the decoupled loadflow equations, active power transfer between two nodes is strongly dependent on the phase difference between the two nodes. An error of even 1 millisecond in synchronization could potentially create a phase difference of about 20 degrees, leading to a large deviation in state estimation.

Such a timing attack can be executed by using GPS spoofing.

4.4.1 GPS Spoofing Attacks

A GPS spoofing attack is an attempt by a malicious party to deceive a GPS receiver to cause it to estimate its position to be other than the correct one, or to estimate the current time to be different than reality, or any combination of the two, by broadcasting counterfeit GPS signals. A common form of attack, termed as a carry-off attack, begins by broadcasting the equivalent of genuine signals. The power of the counterfeit signals is then slowly increased to drown out the real GPS signals — a not-impossible task, given the weakness of GPS signals. Once the receiver is latched on to the counterfeit signal, the signals are slowly changed to induce the receiver away from correct estimates of time and/or position.

While it has been claimed that the capture of the Lockheed RQ-170 drone aircraft in northeastern Iran in December 2011 was an instance of such a carry-off attack [42], and such attacks have been proposed in the academic community, no known example of a malicious spoofing attack has yet been confirmed [43].

A proof-of-concept GPS spoofing attack was demonstrated by Todd Humphreys et al in 2013, using equipment worth 3000 USD to spoof and hijack the multi-million dollar yacht "White Rose" off the coast of Italy.[3]

[3] http://www.engr.utexas.edu/features/superyacht-gps-spoofing/

4.4.2 Attack Methodology

In the electricity market system described above, there are no synchrophasor measurements. To simulate a more restricted attack on a realistic grid, it is assumed that approximately one-third of the buses have PMUs installed, which measure the magnitudes and phases of the voltage and current injections at each bus, from each line. A timing attack on one such PMU will therefore cause a deviation in phase in all the voltage measurements at the bus and current measurements to and from the bus.

A malicious third party wants to attack the system and make a profit from the market, by compromising a number of sensors and sending bogus measurements to the RTO. The attacker is assumed to have the following capabilities:

1. The attacker has full knowledge the underlying system topology.

2. The attacker knows the optimal states Pg^*, Ld^*, and F^* published by the RTO from the ex-ante market.

3. The attacker compromises several subsets of sensors and can manipulate their readings arbitrarily. The attacker can choose which sensor subset to compromise, however due to limited resources, he can only compromise no more than l sensors. Let $\Gamma = \operatorname{diag}(\gamma_1, \gamma_2, \ldots, \gamma_{I+J+L})$, where γ_i is a binary variable that is one if and only if sensor i is compromised by the attacker.

4. The attacker knows the bad data detection algorithm and can defeat it

The bias introduced by the attacker is given by $z^a \in \operatorname{span}(\Gamma)$. Thus, the state estimation Equation (4.6) can be rewritten as

$$\hat{x}' = Pz' = \hat{x} + Pz^a. \tag{4.12}$$

Thus, the residue of Equation (4.7) can be written as

$$r' = r + (I - CP)\, z^a. \tag{4.13}$$

By Triangle inequality,

$$\|r'\|_2 \leq \|r\|_2 + \|(I - CP)\, z^a\|_2\,. \tag{4.14}$$

If $\|(I - CP)\, z^a\|_2$ is small, tending to 0, then the detector will not be able to distinguish between the attacked and unattacked residues. This leads to the definition introduced in [10]:

Definition 7. *The attacker's input z^a is called ϵ-feasible if $\|(I - CP)\, z^a\|_2 \leq \epsilon$.*

The attacker will choose to buy power at bus i and sell it at bus j, and then carry out the attack. In this scenario, his profit per unit power (p) will be the induced change in the nodal price at buses i and j due to the attack:

$$p = \lambda_i - \lambda_i^{\mathrm{DA}} - \lambda_j + \lambda_j^{\mathrm{DA}}, \tag{4.15}$$

where λ^{DA} denotes the day-ahead price at each bus. Using Equation (4.11), it can be seen that the profit as a function of z' will be dependent on the shadow prices as a function of z', $eta\left(z'\right)$ and $\zeta\left(z'\right)$:

$$p\left(z'\right) = \left(H_i - H_j\right)^T \left(\eta\left(z'\right) - \zeta\left(z'\right)\right) - \lambda_i^{\mathrm{DA}} + \lambda_j^{\mathrm{DA}}. \tag{4.16}$$

If L_+ and L_- are defined as:

$$L_+ = \left\{l \middle| H_{l,i} > H_{l,j}\right\}, \tag{4.17}$$

$$L_+ = \left\{l \middle| H_{l,i} < H_{l,j}\right\}, \tag{4.18}$$

$$\begin{aligned} p\left(z'\right) = &\sum_{l \in L_+} \left(H_{l,i} - H_{l,j}\right)^T \left(\eta_l\left(z'\right) - \zeta_l\left(z'\right)\right) \\ &\sum_{l \in L_-} \left(H_{l,j} - H_{l,i}\right)^T \left(\eta_l\left(z'\right) - \zeta_l\left(z'\right)\right) \\ &- \lambda_i^{\mathrm{DA}} + \lambda_j^{\mathrm{DA}}. \end{aligned} \tag{4.19}$$

Thus $p\left(z'\right) > 0$ if $\lambda_j^{\mathrm{DA}} > \lambda_i^{\mathrm{DA}}$, $\hat{F}_l' >_l^F$ min for $l \in L_+$ and $\hat{F}_l' <_l^F$ max for $l \in L_-$.

This leads to the second definition introduced in [10]:

Definition 8. *The attacker's input z^a is called δ-profitable if*

$$E\left[\hat{F}_l'\right] \geq F_l^{min} + \delta,\ \forall l \in L_+$$
$$E\left[\hat{F}_l'\right] \leq F_l^{max} - \delta,\ \forall l \in L_-,$$

where $E\left[\hat{F}'\right] = F^* + Pz^a$.

Hence, the attacker's strategy during the run time is to find an ϵ-feasible z^a such that the margin δ is maximized:

$$\begin{aligned} \underset{z^a \in \text{span}(\Gamma)}{\text{maximize}} \quad & \delta \\ \text{subject to} \quad & \left\|(I - CP)\, z^a\right\|_2 \leq \epsilon \\ & E\left[\hat{F}_l'\right] \geq F_l^{\min} + \delta,\ \forall l \in L_+ \\ & E\left[\hat{F}_l'\right] \leq F_l^{\max} - \delta,\ \forall l \in L_- \\ & \delta > 0. \end{aligned} \tag{4.20}$$

4.5 Simulation

The system used for simulation is the IEEE benchmark 14-bus system, shown in Figure 4.1.

Buses 2, 6, 7, and 9 are assumed to have PMUs installed. The attacker is assumed to have chosen to buy electricity at bus 2 and sell it at bus 4, and solves his convex optimization problem 4.20 to design timing attack

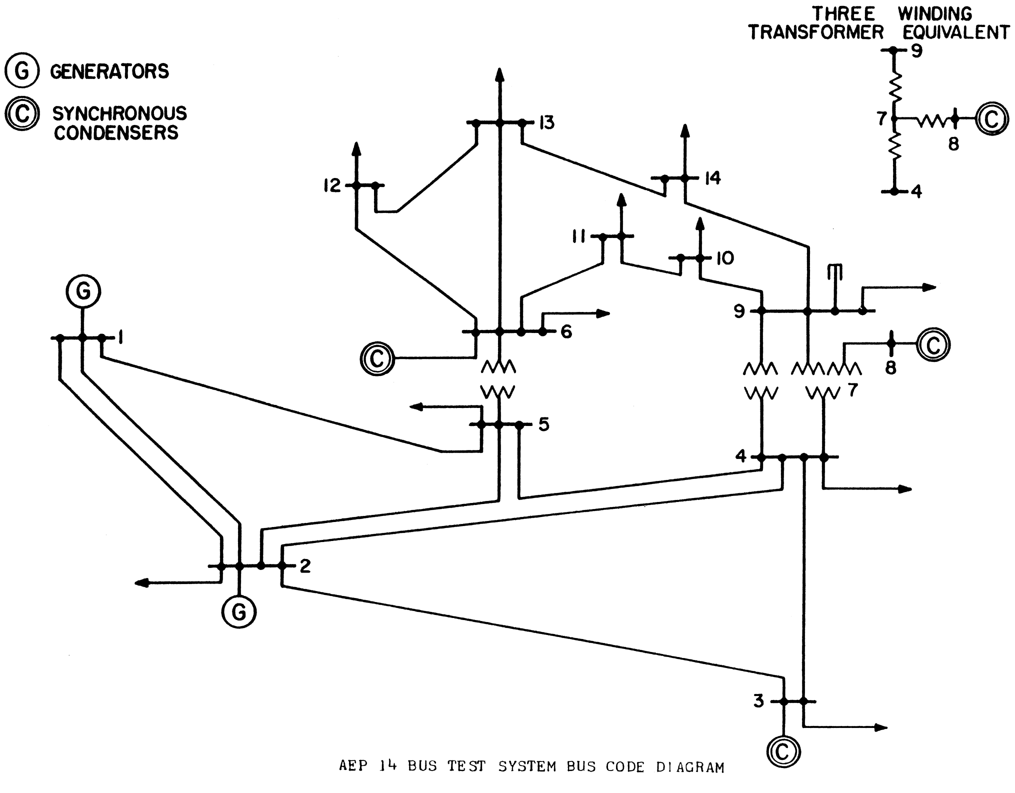

G GENERATORS
C SYNCHRONOUS CONDENSERS
THREE WINDING TRANSFORMER EQUIVALENT
9
7
8
4
13
12
14
11
10
9
6
1
8
7
5
4
2
3
AEP 14 BUS TEST SYSTEM BUS CODE DIAGRAM

Figure 4.1: IEEE 14-Bus System

for the PMU. The attacker restricts his attack to a single PMU, with γ_is corresponding to the affected voltage and current measurements being 1.

The simulation run by Xie et al ([10]) was re-run with these modifications. Figure 4.2 shows the prices at each bus with (red +s) and without (blue ×s) attack. By only attacking one PMU out of 4, i.e., only one bus out of 14, the attacker managed to cause a pricing differential as shown in Figure 4.3.

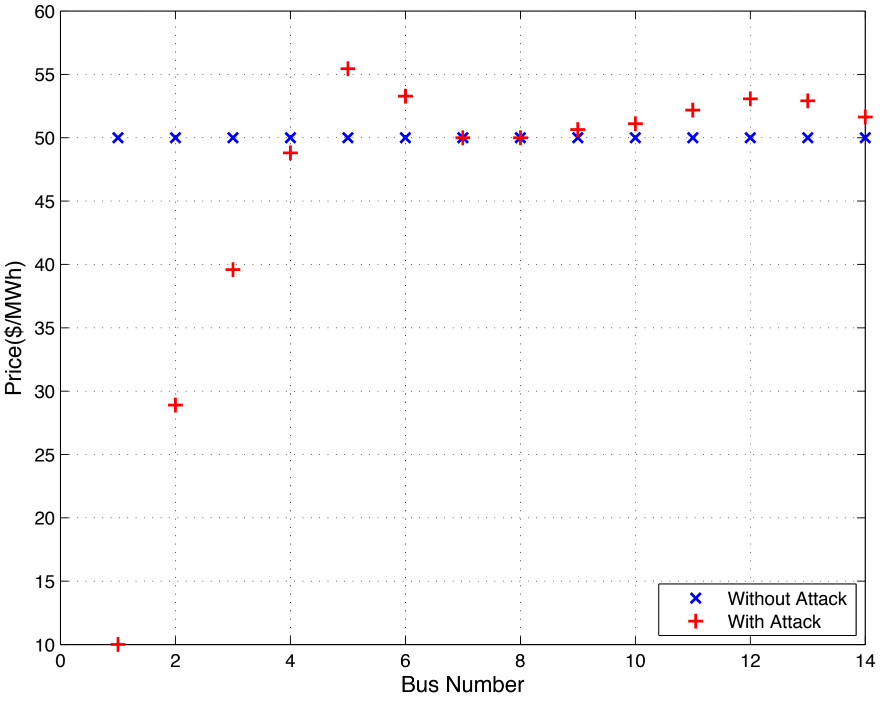


Figure 4.2: Ex-Post Electricity Price at each bus, with (red +s) and without (blue ×s) attack

If the attacker has prior knowledge of his ability to execute the attack, he can outbid his competition in the ex-ante market, and carry out the attack

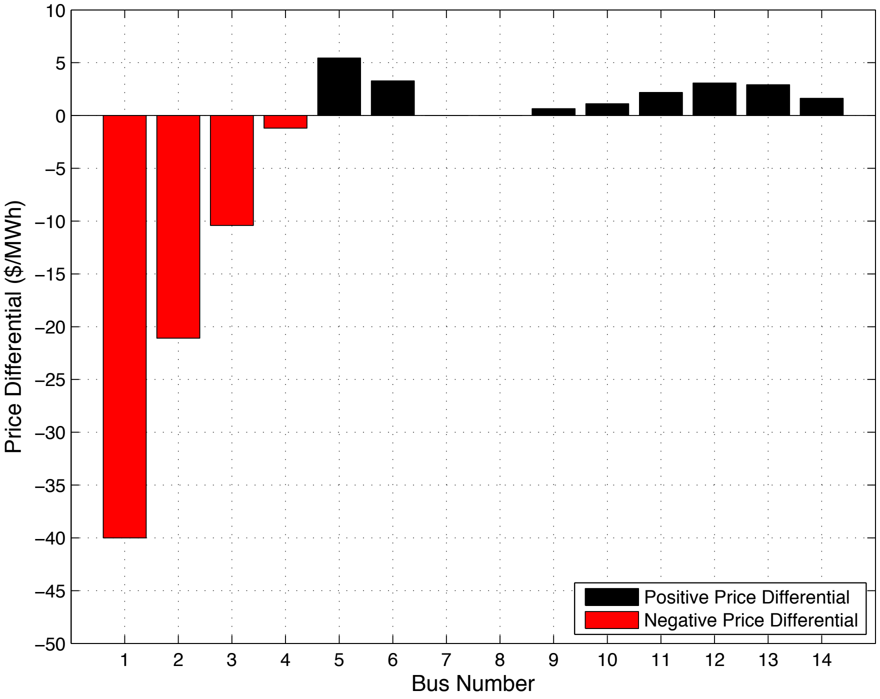


Figure 4.3: Pricing Differential caused at each bus due to attack on one PMU

in real time, thereby affecting the state estimation. In the ex-post market, the attack will cause the buying price for him to fall at bus 2. While the attack also causes the selling price at bus 4 to fall, the overall difference is still profitable to the attacker.

It can be seen that even a very restricted attack scenario, where the attacker can only change the phase measurements at one bus out of fourteen, gives rise to a differential pricing at two nodes chosen by the attacker, without being detected by the bad data detectors. In conjunction with virtual bidding, these integrity attacks can lead to consistent financial profit for the attacker. The potential economic gain for the attackers is thus significant even with small number of sensors being compromised by the attackers.

The next chapter focuses on modeling these theoretical attacks.

Chapter 5

Integrity Attacks: Theoretical Problem

This section focuses on applying simplifying steps to the problem of Section 4. The theoretical detection schemes proposed by Mo et al ([44]) are reviewed, and the limitations faced in the practical application of the schemes are discussed. The detection schemes are then simplified to the most basic problem using binary states and a single class of binary sensors. The problem formulation is then extended to two classes of sensors, and the preliminary results are discussed.

5.1 Previous Work

A conventional method of security is using symmetric and asymmetric encryption and decryption to secure the communications. Cryptographic keys are broken and stolen daily, but even if they were secure, an attacker could directly attack the physical environment of the components, without even touching the communication network. There are other methods of approaching CPS security, most of which rely either on the information content of the system (confidentiality, integrity, availability), or on the robustness of controllers and estimation, detection and identification algorithms. The problem with concentrating on the information content is the lack of a system model, which can blind the detector to a wide variety of attacks (for example, lowering electricity bills by bypassing the meter). On the other hand, robust controllers and algorithms tend to assume random, uncoordinated failures, which is hardly the case during an attack.

In this thesis, we look at the problem of secure detection for a system with a binary state and binary sensors. Although a sensor giving out just one bit of information seems too weak at the first glance, it is more than just an interesting case to look at. For systems using a multitude of distributed sensors for detecting a binary state, it is often superfluous to consider continuous readings from all sensors, and in fact, might prove to be infeasible for both sparse and low-powered communication networks, as well as small embedded processors. It is usual on such a platform for the sensors to be programmed to make a decision based on the information they have, and only communicate this decision over the network, reducing the communication overhead. The controller then makes a decision based on these preliminary decisions.

A similar system has been previously studied by Agah et al ([45]), Alpcan and Başar ([19]), Fuchs and Khargonekar ([46]) and later by Vamvoudakis et al ([47]), by formulating the problem as a zero-sum partial information game in which a detector attempts to minimize the probability of error and an attacker attempts to maximize this probability. The optimal policy recommended by the authors in the latter work is a mixed strategy, where the detector chooses between two rules, based on the perceived probability of attack. This policy is dependent on the estimation of this probability of attack, which, for a lot of systems, is not only extremely difficult to analyze and estimate, but might also change widely based on several external factors.

Kodialam and Lakshman ([48]) also modeled intrusion detection as a zero-sum game, albeit between the service provider and the intruder. Other game-theoretical approach to solving the problem have been proposed by Bier et al ([49]), who used the method increasing the attractiveness of some vectors to the attacker, while designating others as unimportant. The chief drawback of game-theoretical approaches is that the final detection output is possibly a mixed strategy, and not a function of the just the inputs. That is, for the same inputs, the detector output can change randomly based on which policy is chosen, a behavior that may be undesirable in many systems.

Seeking a deterministic solution, we consider the behavior of such a system in the presence of a powerful attacker, without looking to estimate a probability that the adversary will attack. We consider an attack model where the adversary can attack up to a certain number of sensors, while remaining undetected. We provide an insight about what it means for an estimator to be robust in such a scenario, using sensors of different specifications. We analyze the robustness of such a detector for various capabilities of the attacker. We then focus on the case where all the sensors are equivalent, or at least, of similar specifications, and provide a procedure for choosing the detector specifications. We also explore the case where the sensors fall into 2 distinct classes, of different specifications — a case that is of special interest for infrastructures that are undergoing modernization, replacing a few sensors at a time with better versions.

Robust detection with minimax have been previously studied by Huber and Strassen ([50], [51]) and Kassam and Poor ([52]), using uncertainty classes and the detector being designed as a naïve-Bayes or Neymann-Pearson detector. The challenge in such an approach is constructing the least favorable distributions in the uncertainty classes, which are the classes that are supposed to be the hardest for the detector to distinguish.

This section extends the results of [44] in the case of binary sensors and binary cases. The problem of finding the sets defined in the paper has been handled, and a procedure has been proposed to construct these sets in specific cases.

5.2 Problem Formulation

Consider a binary random variable X, with distribution

$$X = \begin{cases} 0 & \text{with probability } P_0 \\ 1 & \text{with probability } P_1 \end{cases}, \tag{5.1}$$

where $P_0, P_1 \geq 0$, and $P_0 + P_1 = 1$. Without loss of generality, let $P_1 \geq P_0$.

To detect X, we have available a vector

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} \in \{0,1\}^m \tag{5.2}$$

of m binary sensor measurements, each of which is conditionally independent from the others given X. Let each sensor have a probability of false alarm (α)

$$P\left(y_i = 1 \middle| X = 0\right) = \alpha_i, \tag{5.3}$$

$$P\left(y_i = 0 \middle| X = 0\right) = 1 - \alpha_i, \tag{5.4}$$

$$i = 1, 2, \ldots, m,$$

and probability of detection (β)

$$P\left(y_i = 1 \middle| X = 1\right) = \beta_i, \tag{5.5}$$

$$P\left(y_i = 0 \middle| X = 1\right) = 1 - \beta_i, \tag{5.6}$$

$$i = 1, 2, \ldots, m.$$

If any of the sensors are actually such that $\alpha_i \geq \beta_i$ for some values of i, the measurements provided by those sensors can be inverted before being used, making $\alpha_i \leq \beta_i$. Thus, without a loss of generality, we can consider $\alpha_i \leq \beta_i \; \forall i$.

In the case where there is no attack, a Bayes detection algorithm suffices.

$$P_0 \prod_{i=1}^{m} \alpha_i^{y_i} \left(1 - \alpha_i\right)^{(1-y_i)} \underset{H_0}{\overset{H_1}{\lessgtr}} P_1 \prod_{i=1}^{m} \beta_i^{y_i} \left(1 - \beta_i\right)^{(1-y_i)} \tag{5.7}$$

where $H_0 \equiv \hat{X} = 0$ and $H_1 \equiv \hat{X} = 1$.

5.2.1 Attack Strategy

It is assumed that an attacker wants to increase the probability that the detector makes an error in detecting X. The attacker has the ability to flip up to l of the m sensor measurements, but the detector does not know which of the m measurements have been manipulated. While the detector knows that at most l measurements have been manipulated, the exact number is also unknown to the detector. This means that any detection scheme $\hat{X} = f(y)$ has to rely on the original measurement vector (y) manipulated by the attack vector (y^a)

$$y^c = y \oplus y^a, \tag{5.8}$$

where $y^a \in \{0,1\}^m$, and $\|y^a\| \leq l$. [1] Here $\oplus$ denotes the element-wise exclusive-or operation. By selecting which bits of y^a are 1, the attacker

[1] In this thesis, we are only dealing with binary states and sensor measurements, where both the 0-norm and the 1-norm are equivalent. Hence, for legibility we choose to drop the subscript, with the understanding that it can be either the 0-norm or the 1-norm. Indeed, the norm $\|\cdot\|$ can very well be replaced by $\|\cdot\|_p^p, 0 \leq p < \infty$ *mutatis mutandis*, without affecting any of the results.

chooses which sensors to attack.

5.2.2 Problem

The detection problem is formalized as a minimax problem where one wants to select an optimal detector

$$\hat{X} = f\left(y^c\right) = f\left(y \oplus y^a\right), \tag{5.9}$$

to minimize the probability of error (or maximize the worst-case probability of detection as derived in section 5.3.1).

5.2.3 Attacker Knowledge

To have the detector follow the Kerckhoffs' Principle which states that, a cryptosystem should be secure even if everything about the system, (except, of course, the key), is public knowledge, we assume that the attacker has full knowledge about f, the state of the system X, and all measurements $y_1, y_2, \ldots, y_m$.

5.3 Results

5.3.1 Robustness and Imperturbable Sets

The question arises about defining robustness of a detector under such an attack. Since we are looking to maximize the probability of detection in the worst possible case, we need to look for all such sensor measurements, such that if those are the measurements provided by the sensors, the adversary can *never* affect enough of them to change the detector output.

Given a detection scheme $f(y)$, let Y_0 be defined as the set of true measurements y, for which any attack vector, which follows the above attack strategy, cannot force the estimate of X to be changed from 0 to 1. Similarly, let Y_1 be defined as the set of true measurements y, for which any attack vector, which follows the above attack strategy, cannot force the estimate of X to be changed from 1 to 0. Formally,

$$Y_0 = \left\{ y \middle| f(y \oplus y^a) = 0, \forall y^a \in \{0,1\}^m, \|y^a\| \leq l \right\}, \tag{5.10}$$

$$Y_1 = \left\{ y \middle| f(y \oplus y^a) = 1, \forall y^a \in \{0,1\}^m, \|y^a\| \leq l \right\}. \tag{5.11}$$

Thus, an attacker cannot affect the detection from any measurement that falls in the set $Y_0 \cup Y_1$, which is, in a sense, the "imperturbable set" for the detector.

The number of sensor measurements that fall in $Y_0 \cup Y_1$ is a measure of the robustness of the detector.

Example

Consider f to be a simple voting scheme, where the detection output depends simply on the majority of the sensor values (m can be considered to be odd to break ties). Let $m = 9$, and $l = 2$. Thus,

$$f(y) = \begin{cases} 0 & \text{if } \|y\| \leq 4 \\ 1 & \text{if } \|y\| > 4. \end{cases} \tag{5.12}$$

It is easy to see that $Y_0 = \left\{ y \middle| \|y\| \leq 2 \right\}$. If $\|y\| \leq 2$, and $\|y^a\| \leq 2$, then $\|y \oplus y^a\| \leq 4$, which will force $f(y) = 0$. Similarly, it is easy to see that $Y_1 = \left\{ y \middle| \|y\| \geq 7 \right\}$. If $\|y\| \geq 7$, and $\|y^a\| \leq 2$, then $\|y \oplus y^a\| \geq 5$, which will force $f(y) = 1$. Thus Y_0 and Y_1, are "good sets" for the detector.

Remark 9. *It is important to note that, $Y_0 \cup Y_1 \neq \{0,1\}^m$, except in the case when $l = 0$ (there is no attacker). That is, there will be measurements possible, which are neither in Y_0 nor in Y_1. For these measurements, the attacker can indeed change the output of the detector. In the above example, if the measurement y is such that $3 \leq \|y\| \leq 6$, the attacker can change the detector output to be what he chooses.*

In the presence of an attacker, there will measurement values for which the attacker is able to cause an error. In a worst-case scenario, a malicious attacker will always cause errors. Thus, only the points in Y_0 and Y_1 contribute to the worst-case probability of detection. Consider $X = 0$. The probability of getting measurement $y \in Y_0$ given $X = 0$ (which will assure $f(y \oplus y^a) = 0$, $\forall y^a \in \{0,1\}^m, \|y^a\| \leq l$) is

$$\sum_{y \in Y_0} \left(\prod_{i=1}^{m} \alpha_i^{y_i} \cdot \prod_{i=1}^{m} (1 - \alpha_i)^{(1-y_i)} \right). \tag{5.13}$$

Similarly, the probability of getting measurement $y \in Y_1$ given $X = 1$ (which will assure $f\left(y \oplus y^a\right) = 1$, $\forall y^a \in \{0,1\}^m, \|y^a\| \leq l$) is

$$\sum_{y \in Y_0} \left(\prod_{i=1}^{m} \beta_i^{y_i} \cdot \prod_{i=1}^{m} \left(1 - \beta_i\right)^{(1-y_i)} \right). \tag{5.14}$$

Thus the total worst-case probability of detection (P) is given by

$$\begin{aligned} P = P_0 &\sum_{y \in Y_0} \left(\prod_{i=1}^{m} \alpha_i^{y_i} \cdot \prod_{i=1}^{m} \left(1 - \alpha_i\right)^{(1-y_i)} \right) \\ &+ P_1 \sum_{y \in Y_1} \left(\prod_{i=1}^{m} \beta_i^{y_i} \cdot \prod_{i=1}^{m} \left(1 - \beta_i\right)^{(1-y_i)} \right). \end{aligned} \tag{5.15}$$

Thus the problem of finding the optimal detector can be formally stated as

$$\begin{aligned} \underset{Y_0, Y_1}{\text{maximize}} \quad P_0 &\sum_{y \in Y_0} \left(\prod_{i=1}^{m} \alpha_i^{y_i} \cdot \prod_{i=1}^{m} \left(1 - \alpha_i\right)^{(1-y_i)} \right) \\ &+ P_1 \sum_{y \in Y_1} \left(\prod_{i=1}^{m} \beta_i^{y_i} \cdot \prod_{i=1}^{m} \left(1 - \beta_i\right)^{(1-y_i)} \right), \end{aligned} \tag{5.16}$$

subject to constraints of the problem, which will be formalized in further sections.

5.3.2 No Fewer Than Half The Sensors Attacked ($l \geq \left\lceil \frac{m}{2} \right\rceil$)

Theorem 10. *If $l \geq \left\lceil \frac{m}{2} \right\rceil$, at least one of Y_0 and Y_1 is empty.*

Proof. $l \geq \left\lceil \frac{m}{2} \right\rceil \Rightarrow m - l \leq l$. Suppose both sets are non-empty. Let

$$y^0 = \begin{pmatrix} y_1^0 & y_2^0 & \cdots & y_m^0 \end{pmatrix}^T \in Y_0, \tag{5.17}$$

$$y^1 = \begin{pmatrix} y_1^1 & y_2^1 & \cdots & y_m^1 \end{pmatrix}^T \in Y_1. \tag{5.18}$$

Consider a measurement y,

$$y = \begin{pmatrix} y_1^0 & y_2^0 & \cdots & y_l^0, y_{l+1}^1 & y_{l+2}^1 & \cdots & y_m^1 \end{pmatrix}. \tag{5.19}$$

Now, $y = y^0 \oplus y^a$, i.e., $y^a = y \oplus y^0$. Since the first l values in y^a are definitely zero, $\|y^a\| \leq m - l \leq l$. By the definition of Y_0 (Equation (5.10)), and the fact that $\|y^a\| \leq l$, it can be concluded that $f(y) = 0$. Let $y = y^1 \oplus y'^a$, i.e., $y'^a = y \oplus y^1$. Since the last $m - l$ values in y'^a are definitely zero, $\|y'^a\| \leq l$. Again by the definition of Y_1 (Equation (5.11)), and the fact that $\|y'^a\| \leq l$, it can be concluded that $f(y) = 1$, which contradicts the previous conclusion. Hence, one of the two sets must be empty. □

Remark 11. *If one of the two sets must empty, the other set can, and in general, should, contain all the possible measurements. Essentially, this scheme is equivalent to the detector disregarding the measurements and making a decision based on the prior probabilities P_0 and P_1. Thus, if $l \geq \left\lceil \frac{m}{2} \right\rceil$ and $P_0 > P_1$, the detector should always detect $\hat{X} = 0$, i.e., the set Y_1 is empty and Y_0 contains all possible measurements. Similarly, if $l \geq \left\lceil \frac{m}{2} \right\rceil$ and $P_1 > P_0$, the detector should always detect $\hat{X} = 1$, i.e., the set Y_0 is empty* *and Y_1 contains all possible measurements.*

The conclusion of Theorem 10 is that if more than half the number of sensors are attacked, the detector should throw away all measurements and always give an output based on the *a priori* probabilities, P_0 and P_1.

Thus from this point onwards, we can consider $l \leq \left\lfloor \frac{m}{2} \right\rfloor$.

5.3.3 Fewer Than Half The Sensors Attacked

Define a distance metric d as follows. Given $a \in A$ and $b \in B$,

$$d(a, b) = \|a - b\|, \tag{5.20}$$

$$d(a, B) = \min_{b \in B} \|a - b\|, \tag{5.21}$$

$$\begin{aligned} d(A, B) &= \min_{a \in A} \|a - B\| \\ &= \min_{a \in A, b \in B} \|a - b\|. \end{aligned} \tag{5.22}$$

Lemma 12. *For any Y_0, Y_1 such that $d(Y_0, Y_1) \geq 2l + 1$ the detector f, $d(y, Y_0) \underset{f(y)=0}{\overset{f(y)=1}{\gtrless}} d(y, Y_1)$, Y_0 and Y_1 are imperturbable sets.*

Proof. We only need to prove that $f(y) = 0 \; \forall y \in Y_0$ and $f(y) = 1 \; \forall y \in Y_1$.

Consider $y \in Y_0$. Let $y^c = y \oplus y^a$. Since the attacker can attack at most l measurements, $\|y^a\| \leq l$. Thus, $\|y^c - y\| \leq l$. Since $y \in Y_0$, the distance metric to Y_0 can only be equal to or smaller than the distance to y, i.e., $d(y^c, Y_0) \leq l$. Since $y \in Y_0$, $d(y, Y_1) \geq 2l + 1$. Since $\|y^c - y\| \leq l$, by the triangle inequality, $d(y^c, Y_1) \geq l+1$. Since, $d(y^c, Y_0) \leq l < 2l+1 \leq d(y^c, Y_1)$, $f(y) = 0$ for all $y \in Y_0$.

Similarly, consider $y \in Y_1$. Let $y^c = y \oplus y^a$. Since the attacker can attack at most l measurements, $\|y^a\| \leq l$. Thus, $\|y^c - y\| \leq l$. Since $y \in Y_1$, the distance metric to Y_1 can only be equal to or smaller than the distance to y, i.e., $d\left(y^c, Y_1\right) \leq l$. Since $y \in Y_1$, $d\left(y, Y_0\right) \geq 2l+1$. Since $\|y^c - y\| \leq l$, by the triangle inequality, $d\left(y^c, Y_0\right) \geq l+1$. Since, $d\left(y^c, Y_1\right) \leq l < 2l+1 \leq d\left(y^c, Y_0\right)$, $f\left(y\right) = 1$ for all $y \in Y_1$. □

Remark 13. *An intuitive way to see this result is that since each attacked sensors counteracts the measurement provided by an unattacked sensor, an attack on l out of m sensors essentially means that the detection is carried out using the measurements provided by $m-2l$ sensors. Thus, $\forall y^0 \in Y_0, y^1 \in Y_1$, $\|y^0 - y^1\| \geq 2l+1$. For example, if $m = 9$ and $l = 2$, 2 unattacked sensors will counteract the effect of 2 attacked sensors, leaving the detector to estimate $\hat{X}$ from 5 sensors. Thus $\|y^0 - y^1\| \geq 5$.*

Thus the problem of finding the optimal detector can be formally stated as

$$\begin{aligned} \underset{Y_0,Y_1}{\text{maximize}} \quad & P_0 \sum_{y \in Y_0} \left(\prod_{i=1}^{m} \alpha_i^{y_i} \cdot \prod_{i=1}^{m} \left(1-\alpha_i\right)^{(1-y_i)} \right) \\ & \qquad + P_1 \sum_{y \in Y_1} \left(\prod_{i=1}^{m} \beta_i^{y_i} \cdot \prod_{i=1}^{m} \left(1-\beta_i\right)^{(1-y_i)} \right) \qquad (5.23) \\ \text{subject to} \quad & d\left(Y_0, Y_1\right) \geq 2l+1. \qquad (5.24) \end{aligned}$$

5.3.4 Special Case: $l = \frac{m-1}{2}$

The result of Lemma 12 is reduces to a simple form, for the particular case where m is odd, and $l = \frac{m-1}{2}$.

Corollary 14. *If $l = \frac{m-1}{2}$, $|Y_0| = |Y_1| = 1$. Further, if $Y_0 = \{y^0\}$ and $Y_1 = \{y^1\}$, $y^0 = \bar{y^1}$.*

Proof. In this case, $d(Y_0, Y_1) \geq 2l + 1$. But $2l + 1 = m$ and the distance between two m-dimensional binary vectors can be at most m. Thus, $d(Y_0, Y_1) = m$. Thus, for any $y^0 \in Y_0$ and $y^1 \in Y_1$, $y^0 = \bar{y^1}$. Suppose that there is another $y'^0 \in Y_0$ such that $d(y'^0, y^1) = m$. By the triangle inequality, $d(y'^0, y^0) \leq 0$, i.e., $y'^0 = y^0$. Thus, Y_0 is a singleton set. Similarly it can be proved that Y_1 is also a singleton set. □

Remark 15. *If none of the sensors are "inverted", then the measurement that will form Y_0 is $y_i = 0\ \forall i$ (thus making $Y_1 = \{y | y_i = 1\ \forall i\}$). To put it formally, if $\alpha_i \leq \beta_i\ \forall i$, then $Y_0 = \left\{\begin{pmatrix} 0 & 0 & \cdots & 0 \end{pmatrix}^T\right\}$ and $Y_1 = \left\{\begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}^T\right\}$.*

5.3.5 Complexity Of The Search-Space

The space of all possible measurements is $\{0,1\}^m$, i.e., there are 2^m possible values of y. Each value can be in Y_0, Y_1, or neither, thus giving rise to 3^{2^m} possible ways of designing Y_0 and Y_1, and hence, the detector.

Having said that, once one of the sets, say Y_0, is fixed, it is possible to expand Y_1 for all measurements such that $d(Y_0, Y_1) \geq 2l + 1$ is not violated, by finding all points at a distance $2l + 1$ or more from each point in Y_0, and then taking the intersection of these. Even considering this reduction, there are 2^{2^m} possible ways of fixing Y_0 and Y_1.

This double-exponential behavior of the enumerations makes a brute-force search impractical beyond a very small value of m — computers will run out of memory by $m = 5$. $m = 6$ is intractable.

In the further sections, we will concentrate on reducing the search-space for some oft-encountered cases.

5.3.6 All Sensors are Equivalent

It is unlikely to ever be the case, that each sensor is unlike every other sensors — in a practical application, most, if not all, sensors would have their false alarm and detection rate equal. Even if the performance parameters are not exactly equal, they would be close enough to each other, that the sensors can be assumed to be equivalent:

$$\alpha_i = \alpha, \tag{5.25}$$

$$\beta_i = \beta, \tag{5.26}$$

$$i = 1, 2, \ldots, m.$$

Thus,

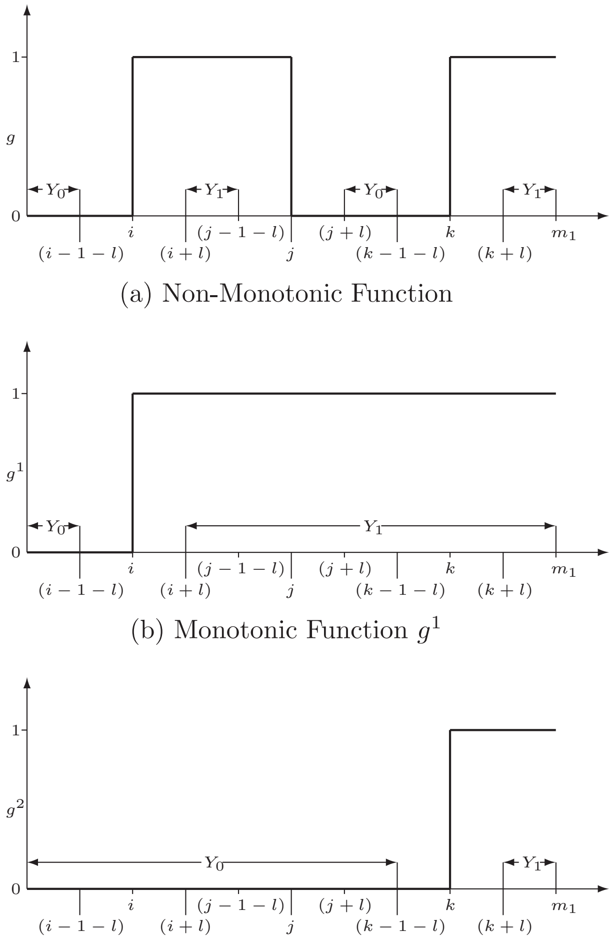
$$P = P_0 \sum_{y \in Y_0} \alpha^{\|y\|} (1-\alpha)^{(m-\|y\|)} + P_1 \sum_{y \in Y_1} \beta^{\|y\|} (1-\beta)^{(m-\|y\|)}. \tag{5.27}$$

The advantage of this assumption lies in the fact that the search for the optimal detector can be confined to only those detector functions that are symmetric in sensor values. Further, for any detector that assumes all sensors are equivalent, the detector function is a symmetric Boolean function, and the output of the detector is a function of only the number of ones or zeros in the measurement y ([53]). Thus, the detector function $f(y)$, where $y = \begin{pmatrix} y_1 & y_2 & \cdots & y_m \end{pmatrix}^T$ can be one of several types of counting functions:

$$\begin{aligned} T_k^n(y) = 1 &\iff \|y\| \geq k && \text{(threshold functions)} \\ E_k^n(y) = 1 &\iff \|y\| = k && \text{(exactly-}k\text{-functions)} \\ C_{k,p}^n(y) = 1 &\iff \|y\| = k \mod p. && \text{(counting functions)} \end{aligned}$$

In this case, however, the optimal detector function, i.e., the function with the maximum worst-case probability of detection (among symmetric Boolean functions) can be proved to be a threshold function, i.e., it is monotonically increasing.

Theorem 16. *The optimal function $g(\|y\|)$, defined to be a symmetric Boolean function with the maximum worst-case probability of detection, is monotonically increasing.*



(a) Non-Monotonic Function

(b) Monotonic Function g^1

(c) Monotonic Function g^2

Figure 5.1: Detector Functions — X-axis is $\|y\|$

Proof. By the assumption that none of the sensors are inverted, $g(0) = 0$ and $g(m) = 1$. Suppose that the function g is not monotonic, and has a "kink". Thus, $\exists i, j, k$, such that $0 \leq i < j < k \leq m_1 \leq m$ and

$$g(n) = \begin{cases} 0 & \text{if } 0 \leq n \leq i-1 \\ 1 & \text{if } i \leq n \leq j-1 \\ 0 & \text{if } j \leq n \leq k-1 \\ 1 & \text{if } k \leq n \leq m_1 \end{cases} \tag{5.28}$$

An example function g with such a "kink" is shown in Figure 5.1a. Each kink in the function can be denoted by unique values of (i, j, k, m_1). In the following argument, we consider only the kink closest to 0.

Since the detector function is given by

$$g(\|y\|) = \begin{cases} 0 & \text{if } d(y, Y_0) > d(y, Y_1) \\ 1 & \text{if } d(y, Y_0) \leq d(y, Y_1), \end{cases} \tag{5.29}$$

where,

$$d(Y_0, Y_1) \geq 2l + 1, \tag{5.30}$$

the subsets of Y_0 and Y_1 that lie in the range $[0, m_1]$ can be computed to be

$$Y_0 = \left\{ y \middle| 0 \leq \|y\| \leq (i-1-l) \right\} \cup \left\{ y \middle| (j+l) \leq \|y\| \leq (k-1-l) \right\} \tag{5.31}$$

$$Y_1 = \left\{ y \middle| (k+l) \leq \|y\| \leq m_1 \right\} \cup \left\{ y \middle| (i+l) \leq \|y\| \leq (j-1-l) \right\} \tag{5.32}$$

Depending upon the value of m_1 as compared to m, there can be other subsets of Y_0 and/or Y_1 beyond the range that we consider. However, the presence of such subsets will not affect the argument.

These sets are also shown in Figure 5.1a. Now consider two other functions, $g^1, g^2 \not\equiv g$ as follows:

$$g^1\left(n\right) = \begin{cases} 0 & \text{if } 0 \le n \le i-1 \\ 1 & \text{if } i \le n \le m_1 \\ g\left(n\right) & \text{if } m_1 \le n \le m \end{cases} \tag{5.33}$$

$$g^2\left(n\right) = \begin{cases} 0 & \text{if } 0 \le n \le k-1 \\ 1 & \text{if } k \le n \le m_1 \\ g\left(n\right) & \text{if } m_1 \le n \le m \end{cases} \tag{5.34}$$

The corresponding subsets of Y_0^1, Y_1^1, Y_0^2, and Y_1^2 within the range $[0, m_1]$ are given by

$$Y_0^1 = \left\{ y \middle| 0 \le \|y\| \le (i-1-l) \right\} \tag{5.35}$$

$$Y_1^1 = \left\{ y \middle| (i+l) \le \|y\| \le m_1 \right\} \tag{5.36}$$

$$Y_0^2 = \left\{ y \middle| 0 \le \|y\| \le (k-1-l) \right\} \tag{5.37}$$

$$Y_1^2 = \left\{ y \middle| (k+l) \le \|y\| \le m_1 \right\} \tag{5.38}$$

These two functions, along with the sets are shown in Figs. 5.1b and 5.1c. It can be seen that g^1 and g^2 are defined in a way to have only one of the two $0 \to 1$ transitions of the first kink in g. Now, using the definition of the worst-case probability of detection, the probability P_d for the detector function g can be given by

$$\begin{aligned} P_d = P_0 \sum_{n=0}^{i-1-l} \alpha^n (1-\alpha)^{m-n} + P_0 \sum_{n=j+l}^{k-1-l} \alpha^n (1-\alpha)^{m-n} \\ + P_1 \sum_{n=i+l}^{j-1-l} \beta^n (1-\beta)^{m-n} + P_1 \sum_{n=k+l}^{m_1} \beta^n (1-\beta)^{m-n} \\ + P_{(m_1,m)}, \end{aligned}$$

where $P_{(m_1,m)}$ denotes the contribution to the worst-case probability of detection, of the part of the function that lies beyond the range $[0, m_1]$ that we consider. Comparatively, the worst-case detection probabilities P_d^1 and P_d^2 for the constructed functions g^1 and g^2 respectively, can be calculated to be

$$\begin{aligned} P_d^1 = P_d - \underbrace{\left(P_0 \sum_{n=j+l}^{k-1-l} \alpha^n (1-\alpha)^{m-n} - P_1 \sum_{n=j+l}^{k-1-l} \beta^n (1-\beta)^{m-n} \right)}_{P_{\text{diff}}} \\ \underbrace{+ P_1 \sum_{n=j-1-l}^{j+l} \beta^n (1-\beta)^{m-n} + P_1 \sum_{n=k-1-l}^{k+l} \beta^n (1-\beta)^{m-n},}_{P_\beta} \end{aligned}$$

and

$$P_d^2 = P_d + \underbrace{\left(P_0 \sum_{n=j+l}^{k-1-l} \alpha^n (1-\alpha)^{m-n} - P_1 \sum_{n=j+l}^{k-1-l} \beta^n (1-\beta)^{m-n} \right)}_{P_{\text{diff}}}$$
$$+ \underbrace{P_0 \sum_{n=i-1-l}^{i+l} \alpha^n (1-\alpha)^{m-n} + P_0 \sum_{n=j-1-l}^{j+l} \alpha^n (1-\alpha)^{m-n}}_{P_\alpha}.$$

That is,

$$P_d^1 = P_d - P_{\text{diff}} + P_\beta$$
$$P_d^2 = P_d + P_{\text{diff}} + P_\alpha.$$

We know that $P_\alpha, P_\beta \geq 0$. Now, for g to be optimal, $P_d \geq P_d^1$ and $P_d \geq P_d^2$. But,

$$\begin{aligned} & P_d \geq P_d^1 \\ \iff & P_d \geq P_d - P_{\text{diff}} + P_\beta \\ \iff & P_{\text{diff}} \geq P_\beta \\ \Rightarrow & P_{\text{diff}} \geq 0, \end{aligned} \tag{5.39}$$

and

$$\begin{aligned} P_d &\geq P_d^2 \\ \iff P_d &\geq P_d + P_{\text{diff}} + P_\alpha \\ \iff -P_{\text{diff}} &\geq P_\alpha \\ \Rightarrow P_{\text{diff}} &\leq 0. \end{aligned} \tag{5.40}$$

The only way these inequalities are satisfied, is if $P_{\text{diff}} = P_\alpha = P_\beta = 0$. This will be the case if $\alpha = \beta$ (in which case, all three probabilities are equal), or $i = j = k$ (there is no kink). The first case is discounted by the assumption that $\alpha < \beta$, and in the second case, all three functions g, g^1, and g^2 are equivalent, which is discounted by the assumption $g^1, g^2 \not\equiv g$. This is a contradiction.

Thus, the worst-case probability of detection of any function g can only be increased by removing the first such kink in g. If the function g has more than one kink, upon removal of the first kink in g, there will be a new "first kink" in the new function. However, the above result can be applied successively to each such kink, leading to the conclusion that the optimal g, the one that has the maximum worst-case probability of detection, has no such kinks, i.e., the optimal g has to be monotonically increasing. □

Since the optimal detector function has only one $0 \rightarrow 1$ transition, it can be defined only by one parameter, the threshold. The results of Lemma 12 can be combined with Theorem 16, to obtain the conditions for the threshold:

Corollary 17. *In a system where all m sensors have equivalent specifications, and the attacker can attack up to l sensors, the sets Y_0 and Y_1 which maximize the worst-case probability of detection such that $d(Y_0, Y_1) \geq 2l+1$, are given by*

$$Y_0 = \left\{ y \middle| \, \|y\| \leq n \right\}, \tag{5.41}$$

$$Y_1 = \left\{ y \middle| \, \|y\| \geq n + 2l + 1 \right\}, \tag{5.42}$$

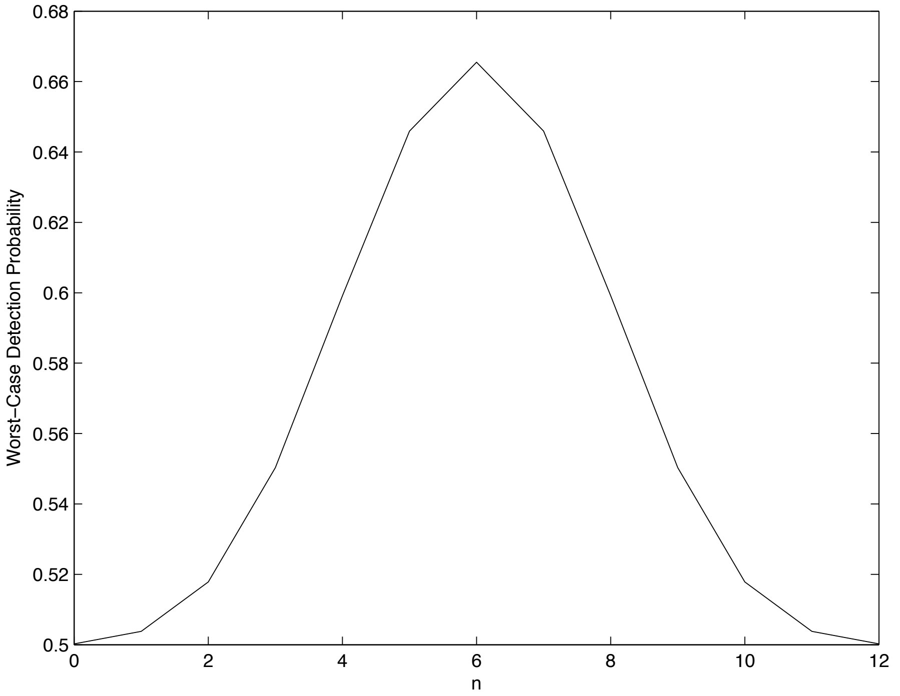
for some integer n such that $0 \leq n \leq \frac{m-1}{2}$. The detector function is therefore given by

$$f(\|y\|) = \begin{cases} 0 & \text{if } \|y\| \leq n + l \\ 1 & \text{if } \|y\| \geq n + l + 1. \end{cases} \tag{5.43}$$

5.3.7 General Values Of l

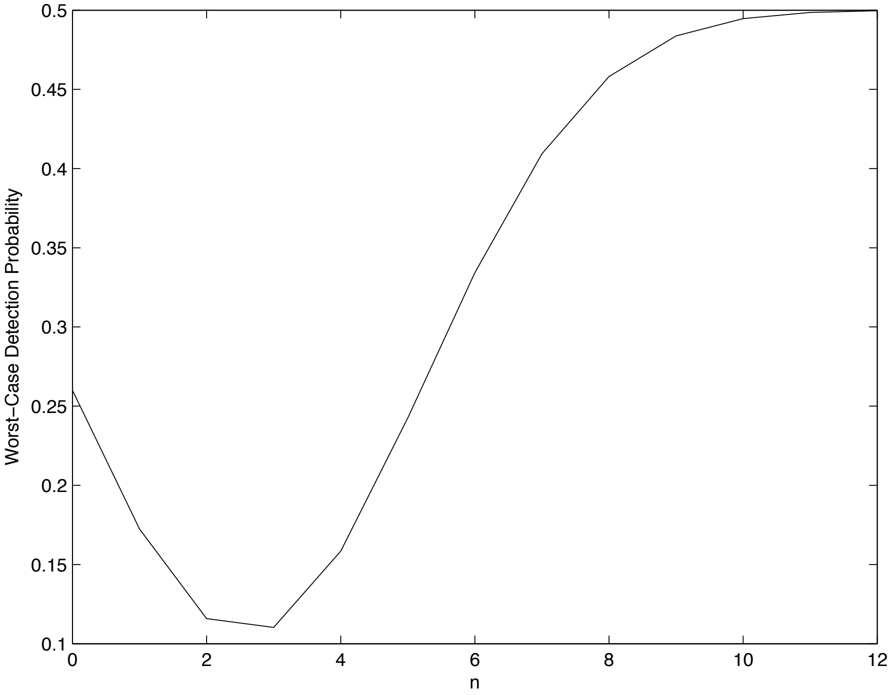
We now consider other values of $l < \left\lfloor \frac{m-1}{2} \right\rfloor$. For given m and l, the worst-case probability of detection P is a function of n parametrically dependent on P_0, P_1, α and β. The shape of the function varies widely with a small change in these values, and cannot be said to be either convex or concave. For example, for $m = 9$ and $l = 3$ we get the plots of worst-case probability of detection P vs. n for different values of α and β, shown in Figure 5.2.

As a result, it is impossible to predict a closed form expression for n. The only solution is to do on exhaustive search for $n = 0$ through $n = m - 2l - 1$. This is a linear search and thus tractable even for large values of m and l.



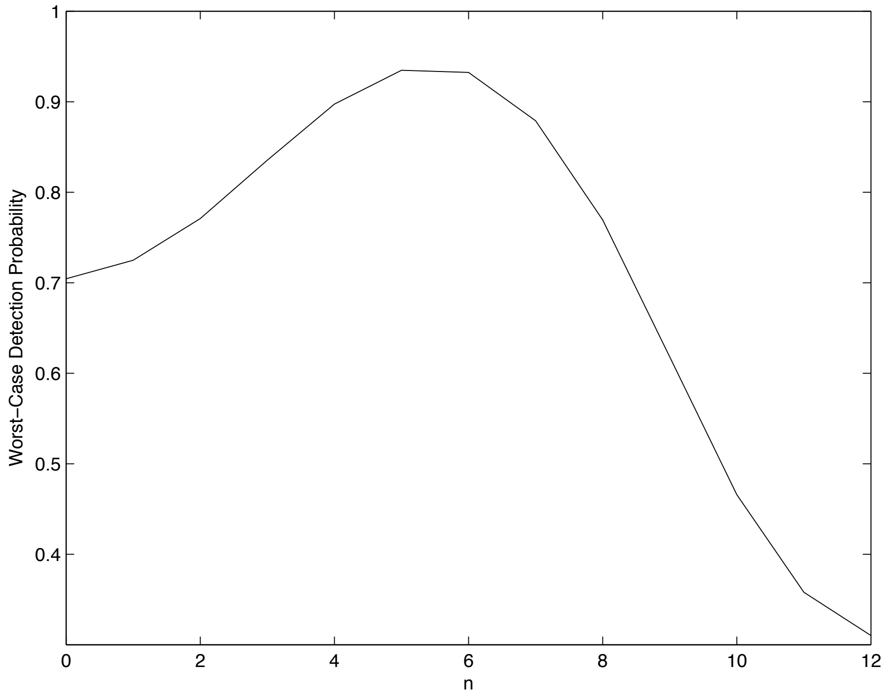
(a) $P_0 = P_1 = 0.5, \alpha = 0.3, \beta = 0.7$

Figure 5.2: Worst-case probability of detection P as a function of n, for $m = 9$ and $l = 3$.



(b) $P_0 = P_1 = 0.5, \alpha = 0.3, \beta = 0.35$

Figure 5.2: Worst-case probability of detection P as a function of n, for $m = 9$ and $l = 3$.



(c) $P_0 = 0.3, P_1 = 0.7, \alpha = 0.2, \beta = 0.8$

Figure 5.2: Worst-case probability of detection P as a function of n, for $m = 9$ and $l = 3$.

5.3.8 Two Classes of Sensors

There is an often-encountered case in practical applications, where the sensors can be grouped into two classes — "good" sensors, and "better" sensors. This is usually the case when the sensors of a legacy network are being upgraded in steps, or when the better sensors are much more expensive than the good ones to be considered worth it. In such a case, a compromise can be reached by only installing a few better sensors, while most of the network is composed of the cheaper sensors. For example, Phasor Measurement Units (PMUs) are so expensive compared to power meters, that only a few substations have them installed. Although the power grid can be considered to be in the process of being upgraded, even the best-case distribution of the PMUs is expected to be around 30% of the total sensors.

$$\alpha_i = \alpha_a, \tag{5.44}$$

$$\beta_i = \beta_a, \tag{5.45}$$

$$i = 1, 2, \ldots, m_a.$$

$$\alpha_i = \alpha_b, \tag{5.46}$$

$$\beta_i = \beta_b, \tag{5.47}$$

$$i = m_a + 1, m_a + 2, \ldots, m_a + m_b = m.$$

Let

$$y = \left(\underbrace{\begin{pmatrix} y_1 & y_2 & \cdots & y_{m_a} \end{pmatrix}}_{y_a} \underbrace{\begin{pmatrix} y_{m_a+1} & y_{m_a+2} & \cdots & y_{m=m_a+m_b} \end{pmatrix}}_{y_b}\right) \tag{5.48}$$

The search for the optimal detector can be confined to only those detector functions that are symmetric in y_a and y_b, making $f\left(y_1, y_2, \ldots, y_m\right) = g\left(\|y_a\|, \|y_b\|\right)$.

$$\begin{aligned} P = P_0 & \sum_{\begin{pmatrix} y_a & y_b \end{pmatrix}^T \in Y_0} \left(\alpha_a^{\|y_a\|}\left(1-\alpha_a\right)^{\left(m_a-\|y_a\|\right)} \cdot \alpha_b^{\|y_b\|}\left(1-\alpha_b\right)^{\left(m_b-\|y_b\|\right)}\right) \\ & + P_1 \sum_{\begin{pmatrix} y_a & y_b \end{pmatrix}^T \in Y_1} \left(\beta_a^{\|y_a\|}\left(1-\beta_a\right)^{\left(m_a-\|y_a\|\right)} \cdot \beta_b^{\|y_b\|}\left(1-\beta_b\right)^{\left(m_b-\|y_b\|\right)}\right). \end{aligned} \tag{5.49}$$

This case reduces to a search over a 2-D space. However, equivalent conditions of monotonicity do not hold. As a counterexample, consider $m_a = 4$, $m_b = 3$, with $P_0 = P_1 = 0.5$ and $\alpha_a = 0.1$, $\beta_a = 0.9$, $\alpha_b = 0.2$, $\beta_b = 0.8$. The optimal Y_0 and Y_1 are given in Figure 5.3.

Thus, the search needs to be carried over a space of $2^{(m_a+1)(m_b+1)}$ possible combinations of Y_0 and Y_1. This is a significant reduction in complexity over the double-exponential nature of the original problem, and tractable for values of $m \leq 12$.

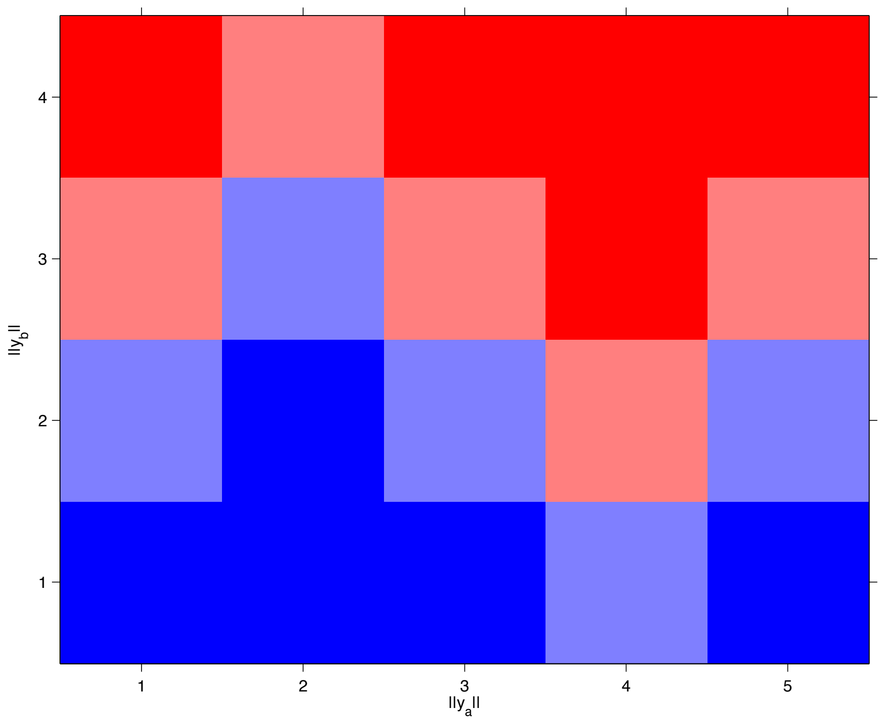


Figure 5.3: Optimal Y_0 (blue) and Y_1 (red) for $m_a = 4$, $m_b = 3$, with $P_0 = P_1 = 0.5$ and $\alpha_a = 0.1$, $\beta_a = 0.9$, $\alpha_b = 0.2$, $\beta_b = 0.8$. The paler colors denote the corresponding decision when the point is neither in Y_0 nor Y_1.

5.4 Correlated Sensors

In cyberphysical systems, the sensors in question monitor a physical system — a system that is constrained to obey physical laws. In such a case, the physical quantities measured by all sensors can scarcely be independent of each other. The measurements of and the noise in each sensor will be correlated to the sensors close to it. This section focuses on modeling the correlation between the sensors and its ramifications on the worst-case probability of detection.

5.5 Correlated Binary Variables

Consider a set of m binary sensors $y_1, y_2, \ldots, y_m$ the measurements of which are not independent. Each sensor has probability of false alarm (α)

$$P\left(y_i = 1 \middle| X = 0\right) = \alpha_i, \tag{5.50}$$

$$P\left(y_i = 0 \middle| X = 0\right) = 1 - \alpha_i, \tag{5.51}$$

$$i = 1, 2, \ldots, m,$$

and probability of detection (β)

$$P\left(y_i = 1 \middle| X = 1\right) = \beta_i, \tag{5.52}$$

$$P\left(y_i = 0 \middle| X = 1\right) = 1 - \beta_i, \tag{5.53}$$

$$i = 1, 2, \ldots, m.$$

It is safe to assume that the correlation coefficient between the sensors is constant, irrespective of the state of the system (X). Even if this weren't true, the correlation of the sensor measurements can be considered separately when $X = 1$ and $X = 0$. Since the derivations are similar, for cleanliness of notation during the rest of the section, the value of the state X will not be specified. The probabilities will instead be denoted as

$$P(y_i = 1) = p_i, \tag{5.54}$$

$$P(y_i = 0) = 1 - p_i, \tag{5.55}$$

$$i = 1, 2, \ldots, m.$$

with the understanding that, if $X = 1$, $p_i = \beta_i$ and if $X = 0$, $p_i = \alpha_i$ for all $i = 1, 2, \ldots, m$.

Now, the probabilities p_i need not be independent. That is, for some $1 \leq i_1 < i_2 \leq m$, $E[y_{i_1} y_{i_2}] \neq E[y_{i_1}] E[y_{i_2}]$. In fact, since more than two variables can be interdependent, for some $1 \leq i_1 < i_2 < \ldots < i_k \leq m$, $1 < k \leq m$,

$$E[y_{i_1} y_{i_2} \ldots y_{i_k}] \neq E[y_{i_1}] E[y_{i_2}] \ldots E[y_{i_k}]. \tag{5.56}$$

If

$$w_i = \frac{y_i - p_i}{\sqrt{p_i(1 - p_i)}}, \tag{5.57}$$

The correlation coefficient $r_{i,j}$ can be written as

$$r_{i,j} = \frac{E\left[y_i y_j\right] E\left[\left(1-y_i\right)\left(1-y_j\right)\right] - E\left[y_i\left(1-y_j\right)\right] E\left[\left(1-y_i\right) y_j\right]}{\sqrt{E\left[y_i\right] E\left[1-y_i\right] E\left[y_j\right] E\left[1-y_j\right]}}. \tag{5.58}$$

Using $E\left[y_i\right] = p_i$ and simplifying the expectations,

$$\begin{aligned} r_{i,j} &= \frac{E\left[y_i y_j\right] - p_i p_j}{\sqrt{p_i\left(1-p_i\right) p_j\left(1-p_j\right)}} \\ &= E\left[w_i w_j\right]. \end{aligned} \tag{5.59}$$

Similarly, the higher correlation coefficients can also be calculated as

$$r_{i_1,i_2,\ldots,i_k} = E\left[w_{i_1} w_{i_2} \ldots w_{i_k}\right]. \tag{5.60}$$

As derived by Bahadur ([54]), the joint probability for a measurement vector $Y = (y_1, y_2, \ldots, y_m)$ can then be written as

$$P\left(y_1, y_2, \ldots, y_m\right) = \prod_{i=1}^{m} p_i^{y_i}\left(1-p_i\right)^{1-y_i} h\left(y_1, y_2, \ldots, y_m\right), \tag{5.61}$$

where

$$\begin{aligned} h\left(y_1, y_2, \ldots, y_m\right) = 1 &+ \sum_{j<k} r_{jk} w_j w_k \\ &+ \sum_{j<k<l} r_{jkl} w_j w_k w_l + \ldots \\ &+ r_{12\ldots m} w_1 w_2 \ldots w_m. \end{aligned} \tag{5.62}$$

This is the probability, calculated by substituting α_1 and β_i for p_i, that causes the manifestation of the factor $h\left(y_1, y_2, \ldots, y_m\right)$ in the worst-case probability of detection P of Equation (5.15).

For m greater than 4 or 5, this distribution can become computationally infeasible. One of the assumptions that are usually made (Emrich and Piedmonte, [55]), is that some of the higher order correlation coefficients $r_{jkl\ldots}$ are zero. The problem with this assumption is that since $r_{jkl\ldots}$ need to satisfy linear inequalities determined by the marginal expectations, they are not free to vary over $[-1, 1]$. Thus by assuming $r_{jkl\ldots}$ are zero, the values of h at some measurement vectors might be negative.

Even the values of r_{jk} cannot be freely chosen from $[-1, 1]$. An intuitive way to see this is to consider the correlations as cosines of angles in L^2. For example, consider three sensors y_1, y_2, and y_3. If $r_{12} > 0$ and $r_{23} > 0$, it can be seen that

$$r_{12}r_{23} - \sqrt{1 - r_{12}^2} \cdot \sqrt{1 - r_{23}^2} \leq r_{13} \leq r_{12}r_{23} + \sqrt{1 - r_{12}^2} \cdot \sqrt{1 - r_{23}^2}. \quad (5.63)$$

Thus, if $r_{12}, r_{23} > \frac{1}{\sqrt{2}}$, then $0 < r_{23} \leq 1$ necessarily. If r_{23} were assumed to be zero, it would violate the triangle inequality.

In the next section we propose a method to overcome this problem by using a different assumption.

5.6 Correlation Assumptions

Zero is as arbitrary a value for the correlation coefficient as any. In fact, assuming $r_{jkl\ldots}$ are zero could potentially make $h\left(y_1, y_2, \ldots, y_m\right)$ negative for some values of $y_1, y_2, \ldots, y_m$. In order to avoid this, we propose that the correlation coefficient be set in the following roundabout manner, such that $h\left(y_1, y_2, \ldots, y_m\right)$ is guaranteed to be non-negative.

The key idea is to specify as many values of $r_{jkl\ldots}$ as possible. Several methods have been proposed to generate binary random variables that have the given correlation values — for example, Emrich and Piedmonte ([55]), and Lunn and Davies ([56]). A method that generates random variables of given 2-correlations by using Poisson processes is proposed by Park et al ([57]). By choosing a suitable method of generating these correlated binary random variables, the remaining correlations can be algebraically computed and used instead of using zeros.

For example, consider $m = 3$ with $p_1 = 0.9$, $p_2 = 0.8$, $p_3 = 0.7$, and the 2-correlation coefficients are given as $r_{12} = 0.1$, $r_{13} = 0.5$ and $r_{23} = 0.5$. Given the 2-correlations, the generation method in [57] can be chosen.[2] Applying

[2] The values are chosen to match one of the examples given in [57].

the method, we get

$$z_1 = \mathcal{P}_1 + \mathcal{P}_2 + \mathcal{P}_3 \tag{5.64}$$

$$z_2 = \mathcal{P}_1 \qquad\qquad + \mathcal{P}_4 + \mathcal{P}_5 \tag{5.65}$$

$$z_3 = \mathcal{P}_1 + \mathcal{P}_2 \qquad + \mathcal{P}_4 \qquad + \mathcal{P}_6, \tag{5.66}$$

where

$$\mathcal{P}_1 = \text{Poisson}\left(\theta_1 = 0.0165\right), \tag{5.67}$$

$$\mathcal{P}_2 = \text{Poisson}\left(\theta_2 = 0.0870\right), \tag{5.68}$$

$$\mathcal{P}_3 = \text{Poisson}\left(\theta_3 = 0.0018\right), \tag{5.69}$$

$$\mathcal{P}_4 = \text{Poisson}\left(\theta_4 = 0.1350\right), \tag{5.70}$$

$$\mathcal{P}_5 = \text{Poisson}\left(\theta_5 = 0.0716\right), \tag{5.71}$$

$$\mathcal{P}_6 = \text{Poisson}\left(\theta_6 = 0.1181\right), \tag{5.72}$$

where Poisson (θ) denotes a Poisson process of intensity θ.

The binary random variables y_1, y_2, and y_3 can be generated from z_1, z_2, and z_3:

$$y_i = \begin{cases} 1 & \text{if } z_i = 0 \\ 0 & \text{otherwise.} \end{cases} \tag{5.73}$$

This is the prescribed method for generation of the y_is. However, we don't actually need to generate the y_is to calculate the unspecified coefficients of correlation (in this case, only r_{123}). Using the definition of r_{123} from Equation (5.60),

$$
\begin{aligned}
r_{123} = & \frac{E\left[y_1 y_2 y_3\right] - p_1 p_2 p_3}{\sqrt{p_1 p_2 p_3 \left(1-p_1\right)\left(1-p_2\right)\left(1-p_3\right)}} \\
& - \sqrt{\frac{p_1}{1-p_1}} r_{23} - \sqrt{\frac{p_2}{1-p_2}} r_{13} - \sqrt{\frac{p_3}{1-p_3}} r_{12} \qquad (5.74)
\end{aligned}
$$

The value of $E\left[y_1 y_2 y_3\right]$ can be computed given the forms of y_1, y_2, and y_3. Since $y_1 y_2 y_3 = 1 \iff y_1 = y_2 = y_3 = 1 \iff z_1 = z_2 = z_3 = 0 \iff \mathcal{P}_1 = \mathcal{P}_2 = \ldots = \mathcal{P}_6 = 0$,

$$
E\left[y_1 y_2 y_3\right] = \prod_{i=1}^{l} e^{-\theta_i}. \qquad (5.75)
$$

Performing the computations, $E\left[y_1 y_2 y_3\right] = e^{-0.4300} = 0.6505$, giving $r_{123} = 0.0109$.

Thus, if these processes were to generate y_1, y_2, and y_3, then the value of r_{123} would not be zero. Although assigning the computed value of 0.0109 to r_{123} of our sensors is exactly as arbitrary as assigning 0, the advantage here lies in the fact that as long as the 2-correlations are consistent, the higher correlations will also be consistent, enough to guarantee the non-negativity of $h\left(y_1, y_2, \ldots, y_m\right)$. All that remains is to use the higher correlation values to figure out the worst-case detection probability.

As with the case of uncorrelated sensors, the following section demonstrates the case where all sensors are equivalent, and all 2-correlations are the same.

5.7 All Sensors Equivalent

If all sensors are equivalent,

$$p_i = p,\ 1 \leq i \leq m, \tag{5.76}$$

$$r_{i,j} = r,\ 1 \leq i < j \leq m. \tag{5.77}$$

This uses a special case of the method given in [57]. Using the simplification,

$$z_i = \mathcal{P} + \mathcal{P}_i, \tag{5.78}$$

where

$$\mathcal{P} = \text{Poisson}\left(\mu\right), \tag{5.79}$$

$$\mathcal{P}_i = \text{Poisson}\left(\nu - \mu\right), \tag{5.80}$$

$$\tag{5.81}$$

where $\mu = \log\left(1 + r\frac{1-p}{p}\right)$ and $\nu = -\log p$. Thus, for $1 \leq i_1, i_2, \ldots, i_k \leq m$, where $1 < k \leq m$, simplifying like the example in the last section,

$$E\left[y_{i_1} y_{i_2} \ldots y_{i_k}\right] = \frac{p^{2k-1}}{\left(p + r\left(1-p\right)\right)^{k-1}}. \tag{5.82}$$

Thus, $E\left[y_{i_1}y_{i_2}y_{i_3}\right] = \frac{p^5}{(p+r(1-p))^2}$ can be used to generate the 3-correlations r_3 as

$$r_3 = \frac{\frac{p^5}{(p+r(1-p))^2} - p^3}{p^3\left(1-p\right)^3} - 3r\frac{p}{1-p}. \tag{5.83}$$

These 3-correlations and $E\left[y_{i_1}y_{i_2}y_{i_3}y_{i_4}\right] = \frac{p^7}{(p+r(1-p))^3}$ can be further used to compute r_4, and so on.

The correlation coefficients can be used to generate $h\left(y_1, y_2, \ldots, y_m\right)$ using Equation (5.62), which in turn generates the probabilities using Equation (5.61), which finally manifests in Equation (5.15), to give us the worst case probability of detection for correlated binary sensors. This can be then used to figure out the optimal form of the detector.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this thesis, the problem of integrity attacks on cyberphysical systems was tackled.

A replay attack model on cyber-physical systems was defined, and the performance of the control system under the attack was analyzed. The conditions under which the classical estimation-control-failure detection strategy is not resilient to a replay attack were characterized, and for such systems a technique using a physical watermarking signal was provided to improve detection at the expense of control performance. The relationships between performance loss, detection rate and the strength of the authentication signal, were characterized. A methodology for optimizing the signal was also provided, based on the trade-off between desired detection performance and allowable control performance loss. To illustrate the applicability of the key idea irrespective of the exact estimator, controller, and detector combination, a similar theoretical analysis was carried out on a system implementing a cross-correlator detection scheme. Several different sets of simulations were carried out to verify the theoretical results and illustrate the optimization of the control signal for a chemical plant.

The susceptibility of state estimators in electrical grids to integrity attacks was illustrated by simulating the effect of a GPS spoofing attack on Phasor Measurement Units installed on the IEEE 14-bus system, which caused a malicious change in the electricity market operations for such a grid, causing potential unethical gains for a player in the bidding market. This attack caused significant damage to the integrity of the power market operations while only using a very restricted attack vector — one that only changes a microsecond variation in the synchronicity of phasor measurements using low-cost equipment to spoof signals from GPS satellites. The vulnerability of SCADA systems to integrity attacks suggested the necessity of securing the state estimators against such attacks. In an effort to reduce the combinatorial complexity of a previously-proposed method of designing detectors, the simplifying assumption of binary sensors estimating binary states was made. This assumption, although simplistic, applies to a large class of distributed sensor networks which employ low-cost and low-power sensors in widely dispersed locations to estimate the system state. A methodology for designing a detector in such a system was proposed. To further improve the detector performance, the consideration of correlation between these sensors was proposed, and analyzed.

6.2 Future Work

This section details the future work possible in the area. The future tracks for replay attacks and integrity attacks are detailed seperately in these subsections.

6.2.1 Replay Attacks

In a real-world scenario, several engineering considerations could be employed to improve the proposed designs. One idea is to use a "duty-cycle" for the watermarking signal — the controller will only send the watermarking signal for a percentage of the running time. This would reduce the average performance loss by the same percentage. The attacker potentially has the remaining percentage of the time of one duty-cycle to carry out his replay attack without detection. This non-watermarked time can be designed by making sure that in the worst-case scenario, the security constraints of the plant are not violated, inspite of the attacker trying his best to drive the system into a non-linear region or instability. This would involve calculating the reachability set for the system for different time intervals.

Another possible direction would be to move away from using an authentication signal that is completely random and IID, and instead describe the authentication signal by designing its autocorrelation. While this can potentially reduce the performance loss, if the attacker can find out the value of the authentication signal at one point in time, he is in a position to figure out the future signal. The solution, then, might be to have an authentication signal that is partly IID and partly predictable, with the caveat that if the attacker can find out the authentication signal at one point of time, the probability of detection of the attack reduces thereafter.

Further decrease in performance loss can also be investigated by restricting the abilities of the attacker. For example, if the attacker can only eavesdrop on a subset of the sensors instead of all of them. Future work will also concentrate on extending these techniques to distributed control systems, where several controllers could potentially exchange their secret watermarking signals, in an effort to align them for minimum performance loss. Avenues are also open to consider more sophisticated attack models, such as a combination of a denial-of-service and replay attack, and so on.

6.2.2 Integrity Attacks

While the methodology of estimating the worst-case probability of detection was demonstrated, in truth it involves laborious algebraic manipulations. For higher number of sensors, a Computer Algebra System such as Mathematica or Maple could be used to derive the form of the detector.

The increase in detection rate by considering the effects of correlation will boost the security of distributed sensor networks that employ binary variables. Future work will involve simulating or implementing such a SCADA system in order to demonstrate the effectiveness of the detector, and possibly implementing these methodologies on a simulation or implementation of a power grid. There are several points in a smart grid, apart from the congestion of lines, where a binary sensor is employed — binary variables can be used to denote the states of circuit breakers, the tripping of lines or damage to transformers, or even control actions like applying a capacitor bank for voltage support. A binary detector that is resilient to integrity attacks can be implemented on a simulation or an implementation of such sensors and actuators.

Bibliography

[1] E. A. Lee, "Cyber Physical Systems: Design Challenges," in *Object Oriented Real-Time Distributed Computing (ISORC), 2008 11th IEEE International Symposium on DOI - 10.1109/ISORC.2008.25*. IEEE, 2008, pp. 363–369.

[2] J. Markoff, "A Silent Attack, But Not A Subtle One," *New York Times*, vol. 160, no. 55176, p. 6, Sep. 2010.

[3] N. Falliere, L. Ó Murchú, and E. Chien. (2011) W32. Stuxnet Dossier. [Online]. Available: http://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/w32_stuxnet_dossier.pdf

[4] A. Matrosov, E. Rodionov, D. Harley, and J. Malcho, "Stuxnet under the microscope," *ESET*, 2010.

[5] J. Appelbaum and L. Poitras, "Als Zielobjekt markiert," *Der Spiegel*, vol. 28, pp. 1–3, Jul. 2013.

[6] D. E. Sanger, "Obama Order Sped Up Wave Of Cyberattacks Against Iran," *New York Times*, vol. 161, no. 55789, Jun. 2012.

[7] J. Carlin. (1997, May) A Farewell To Arms. [Online]. Available: http://www.wired.com/wired/archive/5.05/netizen.html

[8] A. Teixeira, D. Pérez, H. Sandberg, and K. H. Johansson, "Attack models and scenarios for networked control systems," in *Proceedings of the 1st international conference on High Confidence Networked Systems.* New York, NY, USA: ACM, 2012, pp. 55–64.

[9] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on*, 2009, pp. 911–918.

[10] L. Xie, Y. Mo, and B. Sinopoli, "Integrity Data Attacks in Power Market Operations," *Smart Grid, IEEE Transactions on DOI - 10.1109/TSG.2010.2096238*, vol. 2, no. 4, pp. 659–666, Dec. 2011.

[11] E. J. Byres and J. Lowe, "the myths and facts behind cyber security risks for industrial control systems," *Proceedings of the VDE Kongress*, vol. 116, 2004.

[12] A. A. Cárdenas, S. Amin, and S. S. Sastry, "Research challenges for the security of control systems," in *Proceedings of the 3rd conference on Hot topics in security.* San Jose, CA: USENIX Association, 2008, pp. 1–6.

[13] ——, “Secure Control: Towards Survivable Cyber-Physical Systems,” in *Distributed Computing Systems Workshops, 2008. ICDCS ’08. 28th International Conference on DOI - 10.1109/ICDCS.Workshops.2008.40*. IEEE, 2008, pp. 495–500.

[14] S. Amin, A. A. Cárdenas, and S. S. Sastry, “Safe and Secure Networked Control Systems under Denial-of-Service Attacks,” in *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 31–45.

[15] B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M. I. Jordan, and S. S. Sastry, “Kalman filtering with intermittent observations,” *Automatic Control, IEEE Transactions on*, vol. 49, no. 9, pp. 1453–1464, 2004.

[16] L. Schenato, B. Sinopoli, M. Franceschetti, K. Poolla, and S. S. Sastry, “Foundations of Control and Estimation Over Lossy Networks,” *Proc. IEEE*, vol. 95, no. 1, pp. 163–187, 2007.

[17] A. S. Willsky, “A Survey of Design Methods for Failure Detection in Dynamic Systems,” *Automatica*, vol. 12, no. 6, pp. 601–611, Nov. 1976.

[18] R. F. Stengel and L. R. Ray, “Stochastic robustness of linear time-invariant control systems,” *Automatic Control, IEEE Transactions on*, vol. 36, no. 1, pp. 82–87, 1991.

[19] T. Alpcan and T. Başar, "A game theoretic approach to decision and analysis in network intrusion detection," in *Decision and Control, 2003. Proceedings. 42nd IEEE Conference on.* IEEE, 2003, pp. 2595–2600 Vol.3.

[20] S. Sundaram and C. N. Hadjicostis, "Structural Controllability and Observability of Linear Systems Over Finite Fields with Applications to Multi-Agent Systems," *Automatic Control, IEEE Transactions on*, vol. 58, no. 1, pp. 60–73, 2013.

[21] L. Lazos and R. Poovendran, "SeRLoc: Robust localization for wireless sensor networks," *ACM Trans. Sen. Netw.*, vol. 1, no. 1, pp. 73–100, 2005.

[22] L. Lazos, R. Poovendran, and S. Čapkun, "ROPE: robust position estimation in wireless sensor networks," *Proceedings of the 4th international symposium on Information processing in sensor networks*, p. 43, 2005.

[23] F. Pasqualetti, A. Bicchi, and F. Bullo, "Distributed intrusion detection for secure consensus computations," in *Decision and Control, 2007 46th IEEE Conference on*, 2007, pp. 5594–5599.

[24] F. Pasqualetti, F. Dörfler, and F. Bullo, "Cyber-physical security via geometric control: Distributed monitoring and malicious attacks," *IEEE Conf. on Decision and Control*, Dec. 2012.

[25] M. Zhu and S. Martínez, "Attack-resilient distributed formation control via online adaptation," in *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on*, 2011, pp. 6624–6629.

[26] A. Giani, S. S. Sastry, K. H. Johansson, and H. Sandberg, "The VIKING project: An initiative on resilient control of power networks," in *Resilient Control Systems, 2009. ISRCS '09. 2nd International Symposium on DOI - 10.1109/ISRCS.2009.5251361.* IEEE, 2009, pp. 31–35.

[27] J. P. Hespanha, P. Naghshtabrizi, and Y. Xu, "A Survey of Recent Results in Networked Control Systems," in *Proceedings of the IEEE*, 2007.

[28] G. Dán and H. Sandberg, "Stealth Attacks and Protection Schemes for State Estimators in Power Systems," in *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*, 2010, pp. 214–219.

[29] H. Sandberg, A. Teixeira, and K. H. Johansson, "On Security Indices for State Estimators in Power Networks," in *First Workshop on Secure Control Systems, Cyber Physical Systems Week 2010*, Apr. 2010.

[30] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure state-estimation for dynamical systems under active adversaries," in *Communication, Control,*

and Computing (Allerton), 2011 49th Annual Allerton Conference on, 2011, pp. 337–344.

[31] ——, "Secure estimation and control for cyber-physical systems under adversarial attacks," *arXiv*, May 2012.

[32] R. A. Maronna, D. R. Martin, and V. J. Yohai, *Robust Statistics*, ser. Theory and Methods. Wiley, Jun. 2006.

[33] P. J. Huber and E. M. Ronchetti, *Robust Statistics*. Wiley, Sep. 2011.

[34] A. Abur and A. G. Expósito, *Power System State Estimation*, ser. Theory and Implementation. CRC Press, Mar. 2004.

[35] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Trans. Inf. Syst. Secur.*, vol. 14, no. 1, pp. 1–33, 2011.

[36] R. K. Mehra and J. Peschon, "An innovations approach to fault detection and diagnosis in dynamic systems," *Automatica*, vol. 7, no. 5, pp. 637–640, Sep. 1971.

[37] P. E. Greenwood and M. S. Nikulin, *A guide to chi-squared testing*. John Wiley & Sons, Apr. 1996.

[38] L. L. Scharf and C. Demeure, *Statistical Signal Processing: Detection, Estimation And Time Series Analysis*. Addison-Wesley Pub. Co., 1991.

[39] J. J. Downs and E. F. Vogel, “A plant-wide industrial process control problem,” *Computers & Chemical Engineering*, vol. 17, no. 3, pp. 245–255, Jan. 1993.

[40] N. L. Ricker, “Model predictive control of a continuous, nonlinear, two-phase reactor,” *Journal of Process Control*, vol. 3, no. 2, pp. 109–123, Sep. 1995.

[41] F. Li, Y. Wei, and S. Adhikari, “Improving an Unjustified Common Practice in Ex Post LMP Calculation,” *Power Systems, IEEE Transactions on*, vol. 25, no. 2, pp. 1195–1197, 2010.

[42] S. Peterson and P. Faramarzi, “Iran hijacked US drone, says Iranian engineer,” *The Christian Science Monitor*, Dec. 2011.

[43] T. E. Humphreys, B. M. Ledvina, M. L. Psiaki, B. W. O’Hanlon, and P. M. Kintner Jr, “Assessing the spoofing threat: Development of a portable GPS civilian spoofer,” *Proceedings of the ION GNSS International Technical Meeting of the Satellite Division*, 2008.

[44] Y. Mo, J. P. Hespanha, and B. Sinopoli, “Robust detection in the presence of integrity attacks,” in *American Control Conference (ACC), 2012 DO* -, 2012, pp. 3541–3546.

[45] A. Agah, S. K. Das, K. Basu, and M. Asadi, “Intrusion detection in sensor networks: a non-cooperative game approach,” in *Network Com-*

puting and Applications, 2004. (NCA 2004). Proceedings. Third IEEE International Symposium on. IEEE, 2004, pp. 343–346.

[46] Z. E. Fuchs and P. P. Khargonekar, "Games, deception, and Jones' Lemma," in *American Control Conference (ACC), 2011 DO -.* IEEE, 2011, pp. 4532–4537.

[47] K. G. Vamvoudakis, J. P. Hespanha, B. Sinopoli, and Y. Mo, "Adversarial detection as a zero-sum game," in *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on*, 2012, pp. 7133–7138.

[48] M. Kodialam and T. V. Lakshman, "Detecting network intrusions via sampling: a game theoretic approach," in *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, 2003, pp. 1880–1889.

[49] V. Bier, S. Oliveros, and L. Samuelson, "Choosing What to Protect: Strategic Defensive Allocation against an Unknown Attacker," *Journal of Public Economic Theory*, vol. 9, no. 4, pp. 563–587, 2007.

[50] P. J. Huber, "A Robust Version of the Probability Ratio Test," *The Annals of Mathematical Statistics*, vol. 36, no. 6, pp. 1753–1758, Jan. 1965.

[51] P. J. Huber and V. Strassen, “Minimax Tests and the Neyman-Pearson Lemma for Capacities,” *The Annals of Statistics*, vol. 1, no. 2, pp. 251–263, Jan. 1973.

[52] S. A. Kassam and H. V. Poor, “Robust techniques for signal processing: A survey,” *Proc. IEEE*, vol. 73, no. 3, pp. 433–481, 1985.

[53] I. Wegener, “The complexity of symmetric boolean functions,” *Computation theory and logic*, 1987.

[54] R. R. Bahadur, “A Representation of the Joint Distribution of Responses to n Dichotomous Items,” in *Studies in Item Analysis and Prediction (Stanford Mathematical Studies in Social Sciences VI)*, H. Solomon, Ed. Stanford University Press, 1961, pp. 158–168.

[55] L. J. Emrich and M. R. Piedmonte, “A Method for Generating High-Dimensional Multivariate Binary Variates,” *The American Statistician*, vol. 45, no. 4, pp. 302–304, Nov. 1991.

[56] A. D. Lunn and S. J. Davies, “A note on generating correlated binary variables,” *Biometrika*, vol. 85, no. 2, pp. 487–490, Jun. 1998.

[57] C. G. Park, T. Park, and D. W. Shin, “A Simple Method for Generating Correlated Binary Variates,” *The American Statistician*, vol. 50, no. 4, pp. 306–310, Nov. 1996.