Optimal Cyber-Attacks under Energy and Stealthiness Constraints in Cyber-Physical Systems

by

Jing Zhou

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Control Systems

Department of Electrical and Computer Engineering
University of Alberta

Abstract

Cyber-physical systems (CPSs) integrate the cyber world and physical entities via seamless combinations of sensing, communication, and control. One typical feature of CPSs is that massive data packets are transmitted through unreliable wireless networks, which can be intercepted or manipulated by malicious agents. These cyber-attacks may lead to confidential information leakage, system performance degradation, and even serious industrial incidents. As a prerequisite, the investigation of worst-case attacks from adversary's perspective is essential to reveal vulnerabilities of CPSs and establish a basis for subsequent development of countermeasures. Therefore, this thesis focuses on the design of worst-case attacks in industrial CPSs with energy and stealthiness constraints.

Two research topics are considered. First, we study the scenario where an adversary launches denial-of-service (DoS) attacks against control channels of a linear quadratic regulator (LQR). Owing to energy constraints, the attacker can only launch consecutive attacks with a fixed horizon to maximize the LQR control cost. Necessary and sufficient conditions are derived under which the optimality of attacking from the initial instant can be preserved despite the randomness of initial states. Second, we consider the scenario that malicious agents can intercept and modify sensor measurements of a remote state estimator, with the purpose to degrade the estimation quality while remaining undetected by anomaly detectors. This scenario is composed of three topics: *i*) An innovation-based linear attack fusing all available information is proposed, which clarifies a counter-intuitive issue in existing work. Explicit expressions of optimal stealthy attack coefficients are obtained without solving optimization problems numerically. *ii*) The optimal information-based attack that achieves the maximum greedy performance and deceives χ^2 detectors is derived. For both attacks with strict and relaxed stealthiness, the optimal compromised innovation is shown to be a linear function of the conditional minimum mean-square error (MMSE) estimate of prediction errors. A unified framework and a separation principle are proposed to handle more general scenarios that the attacker has access to different levels of online information. *iii*) The optimal information-based stealthy attack leading to the maximum holistic performance is obtained. The compromised innovation is constructed as a linear combination of the MMSE estimates of all historical prediction errors; then the combination coefficients are obtained by solving a convex optimization problem. Moreover, the proposed attack can be generalized to deceive interval χ^2 detectors with different lengths. It is shown that the worst-case attack effect is determined by both the amount of online information and the duration of the detection interval.

The effectiveness of the proposed methods is demonstrated by theoretical analysis, numerical examples, as well as comparative studies with existing work. These findings lead us to a better understanding of vulnerabilities of industrial CPSs and facilitate development of protective measures.

Preface

The research work in Chapter 2 was part of an international research collaboration with Dr. Yuzhe Li at Northeastern University, Shenyang, China. The ideas in Chapters 2–6 were from discussions with Prof. Tongwen Chen and Dr. Jun Shang at the University of Alberta. The algorithms, mathematical derivations and proofs, and numerical simulations were my original work, as well as the introduction in Chapter 1.

- Chapter 2 has been published as: Jing Zhou, Jun Shang, Yuzhe Li, and Tongwen Chen, Optimal DoS attack against LQR control channels. *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 68, no. 4, pp. 1348-1352, April 2021.

- A preliminary version of Chapter 3 has been published as: Jing Zhou, Jun Shang, and Tongwen Chen, Optimal linear FDI attacks with side information: A comparative study, in *4th IEEE International Conference on Industrial Cyber-Physical Systems (ICPS)*, Vancouver, May 2021. A full version has been submitted to *IEEE Transactions on Control of Networked Systems*.

- Chapter 4 has been submitted for publication as: Jing Zhou, Jun Shang, and Tongwen Chen, Optimal deception attacks against remote state estimation: An information-based approach, *IEEE Transactions on Automatic Control*. (Accepted without changes as a full paper)

- A preliminary version of Chapter 5 has been published as: Jing Zhou, Jun Shang, and Tongwen Chen, Worst-case stealthy false-data injection attacks on remote state estimation, in *47th Annual Conference of the IEEE Industrial Electronics Society (IECON)*, Toronto, Oct. 2021. A full version has been submitted to *Automatica.* (Provisionally accepted as a regular paper)

- Chapter 6 has been submitted for publication as: Jing Zhou, Jun Shang, and Tongwen Chen, Deception attacks on Kalman filtering with interval estimation performance loss, *5th IFAC Workshop on Linear Parameter Varying Systems*, Montreal, Sept. 2022. (Accepted)

Acknowledgements

I would like to express my sincere gratitude to many people for their help and support. First and foremost, I would like to thank my supervisor, Prof. Tongwen Chen, who admitted me in this doctoral program four years ago when I showed little academic potential. His kind encouragement, guidance, and rigorous research attitude are essential in accomplishing the thesis. I would also like to thank my committee members Drs. Qing Zhao and Mahdi Tavakoli Afshari for their precious time and valuable feedback.

Furthermore, I am deeply thankful to Drs. Jun Shang, Yuzhe Li, and Hao Yu for their valuable discussions, comments, and suggestions. The research on cyber-security would not go smoothly without their kind help. I would like to thank all current and previous members of the research group, Junyi Yang, Md Rezwan Parvez, Mani Hemanth Dhullipalla, Haniyeh Seyed Alinezhad, Li Deng, Harikrishna Rao Mohan Rao, Boyuan Zhou, Wenkai Hu, and Shimin Wang for their kindness, encouragement, and help. In addition, I would like to thank all my badminton friends, Hao Li, Ke Wang, Mengnan Zhu, and Yuzi Liu, who have greatly enriched my campus life in the past four years.

I also wish to acknowledge the financial support from the Chinese Scholarship Council (CSC) and Natural Sciences and Engineering Research Council of Canada (NSERC).

Last but not the least, greatest thanks also go to my family members and friends for their companion mentally or physically over the past years. Again, to people mentioned or not, my thankfulness and gratitude are beyond words and can never be written well enough.

Contents

List of Tables

List of Figures

List of Symbols

$\mathbb{R}$ $(\mathbb{N})$	Set of Real (Natural) Numbers
$\mathbb{S}_{++}^{n}$ $(\mathbb{S}_{+}^{n})$	Set of Positive (Semi-)definite Matrices
$\mathrm{Tr}(X), \lvert X \rvert$	Trace and Determinant of Matrix X
$\lambda_i(X)$	ith Largest Eigenvalue (in Modulus) of X
$\sigma_i(X)$	ith Largest Singular Value of X
$\mathrm{Im}(X), \mathrm{Ker}(X)$	Image and Kernel Spaces of X
$\Re_{\lambda_i}(X)$	Real Part of the ith Eigenvalue of X
X^{+}	Moore-Penrose Pseudoinverse of X
I_n	n by n Identity Matrix
$0_{n\times m}$ (0_n)	n by m (n by n) Matrix with Zero Entries
$\mathrm{blkdiag}\{\cdot\}$	Block Diagonal Matrix
$[\![a, b]\!]$	Set $\{x \in \mathbb{N} \lvert a \leq x \leq b\}$
$\mathcal{N}(\mu, \Sigma)$	Gaussian Distribution with Mean μ and Covariance Σ
$p_X(t)$	Probability Density Function of X
$\mathbb{E}[X], \mathrm{Cov}[X]$	Expectation and Covariance of X
$h(X)$	Lyapunov Operator: $h(X) \triangleq AXA^{\mathrm{T}} + W$
$g_{[C,V]}(X)$	Riccati Operator: $g_{[C,V]}(X) \triangleq X - XC^{\mathrm{T}}(CXC^{\mathrm{T}} + V)^{-1}CX$
$\mathbb{BL}_{[p,q]}(m, n)$	Set of $p \times q$ Block Lower Triangular Matrices Whose Blocks are $m \times n$ (Real) Matrices

List of Acronyms

CPS	Cyber-Physical Systems
DoS	Denial-of-Service
FAR	False Alarm Rate
FDI	False Data Injection
ICC	Input Covariance Constraint
i.i.d.	independent and identically distributed
KLD	Kullback–Leibler Divergence
LMI	Linear Matrix Inequality
LQG	Linear Quadratic Gaussian
LTV	Linear Time-Varying
MMSE	Minimum Mean-Square Error
MTD	Moving Target Defense
SCADA	Supervisory Control and Data Acquisition
SDP	Semi-definite Program
SVD	Singular Value Decomposition
UAV	Unmanned Aerial Vehicle

Chapter 1

Introduction

In this chapter, the research background for cyber-physical system security is introduced and a literature survey is provided to summarize the recent development of worst-case attacks and defensive countermeasures. Thereafter, the contributions of the thesis are listed, followed by a thesis outline.

1.1 Research Background

The last decade witnessed a rapid progress in the development of cyber-physical systems (CPSs), which are tight integrations of computational, networking and physical components. CPSs provide a general modeling framework that covers various industrial processes and critical infrastructures, e.g., power grids [41, 43, 45], water distribution networks [2, 75, 79], intelligent transportation [21, 55], smart medical devices [13, 33], and industrial control systems [1, 32, 80, 82, 91]. The normal operation of CPSs depends on reliable transmission of data packets, which could be craftily manipulated by malicious agents particularly if wireless networks are deployed. *Stuxnet* is one such well-known cyber-worm that caused great damage to nuclear facilities in Iran by injecting falsified commands [32]. In 2015, a synchronized and coordinated cyber-attack compromised three Ukrainian regional power distribution companies, resulting in electric outages affecting approximately 225,000 customers for several hours [41]. A recent cyber-attack that crippled the largest fuel pipeline in the U.S. and led to energy shortages across the east coast was

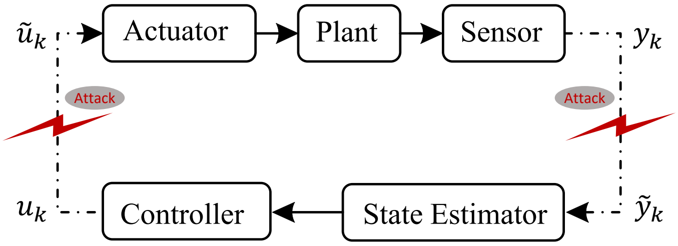

Figure 1.1: Cyber-attacks on industrial control systems.

Security of CPSs can be regarded as a composition of three main attributes: confidentiality, availability, and integrity. The corresponding cyber-threats impairing these properties are respectively termed as eavesdropping attacks, denial-of-service (DoS) attacks, and false-data injection (FDI) attacks [14, 27, 76]. The later two contribute the greatest portion in real-world incidents and have been the focus of academic research on CPS security since a decade ago. Fig 1.1 illustrates typical cyber-attacks targeting an industrial control system, where the dashed lines represent vulnerable data transmission channels. In DoS attacks, the adversary broadcasts noisy data to block communication channels between different components; as a result, the useful information is not available at the receiver's side [6, 81, 89, 90]. FDI attacks, also known as deception attacks, require more resources in practical implementation, since attackers need to intrude into communication links and modify original packets or inject falsified data. In both cases, the nominal performance of CPSs will be significantly damaged; the impact can be increased control costs [82, 91], degraded state estimation quality [4, 24, 89], and even instability of closed-loop systems [43, 51]. Eavesdropping attacks seem moderate for that the attacker's behavior has no direct impact on the system performance, but the leak of critical information can also lead to devastating consequences [17, 25, 88]. Other less common cyber-threats in power distribution grids include topology poisoning, load redistribution, and data framing attacks [47]. Despite inevitability of these malicious attacks, it is impossible for adversaries to launch uncontrolled attacks owing to countermeasures deployed by system defenders, such as the virus firewalls, anomaly detectors, and data encryption mechanisms [49, 60, 65]. The attacker's resource budget and restricted information also confine the feasible set of attack policies.

To enhance the security level of CPSs, it is necessary to study cybersecurity from attacker's perspective, investigating inherent weaknesses of CPSs by exploring worst-case attack strategies. In existing studies, synthesizing optimal attacks that cause the maximum performance loss and developing defensive countermeasures to mitigate their negative impacts are two major research topics.

1.1.1 Worst-Case Cyber-Attacks

The ancient proverb "If you know both the enemy and yourself, you will fight a hundred battles without danger of defeat" points out the necessity and significance to study optimal cyber-attacks that maximize the adversary's benefit. Using "optimal", we emphasize attack impacts from attackers' perspective. It is also called "worst-case" attack when we highlight system performance degradation from defenders' point of view. In an adversarial network environment, investigating the system performance evolution under influence of malicious attacks is a prerequisite for subsequent development of defensive countermeasures [27].

Launching cyber-attacks is a resource-consuming task. In DoS attacks that block wireless channels, the adversary may have to configure powerful emitters to mask original signals [48]. The attacker should design a sophisticated DoS schedule to achieve the maximum attack impact with limited energy supply. For example, in the case that an attacker intends to compromise a remote state estimator but can block only limited steps of measurement transmission in a fixed horizon, there exists an optimal schedule that causes the greatest estimation performance loss [89]. Unlike DoS attacks that can be readily noticed by defenders, FDI attacks are carefully synthesized to cause severe consequences without triggering anomaly detectors. The capacity to avoid being detected is called stealthiness, which is the main property that makes FDI attacks differ from randomly occurred component faults. The pioneering work on stealthy FDI attacks dates back to [45] in smart grids and [51] in dynamic linear systems. For DoS attacks, the optimal trade-off between attack effects and energy consumption is the main concern; while in FDI attacks more efforts have been paid to reach a satisfactory balance between performance loss and stealthiness level.

Additionally, attack performance depends also on the amount of information the adversary can gain [24, 97, 98]. When synthesizing optimal attacks, we usually assume the worst happens, i.e., a malicious agent has access to all necessary resources and information to facilitate his/her purpose. The information includes system parameters, controller and estimator configurations, and online data. This assumption is in accordance with the Kerckhoffs's principle, which states that the security of a system should not rely on its obscurity [73]. Though difficult, it is often assumed that adversaries can infer system parameters using techniques like system identification and controller invasion [16, 46]. The online information refers to the eavesdropped data in unreliable transmission links, namely, real-time sensor measurements and control commands. In some cases, an attacker can not only intercept original sensor outputs but also obtain more information of system states by placing extra sensors [24, 40, 97]. How to fully utilize all available information to achieve the maximum attack impact is a challenge task that has not been fully addressed in existing studies.

As a summary, the design of worst-case attacks is to synthesize optimal attack sequences that cause the maximum performance loss in CPSs, with the following practical concerns:

1. The available information for attackers should be fully utilized.

2. The consumed attack energy must not exceed the budget limit.

3. The attack is capable of deceiving anomaly detectors.

1.1.2 Defensive Countermeasures

The ultimate goal of current studies is to develop efficient countermeasures that eliminate or mitigate the impacts of cyber-threats. These protective mechanisms can be classified as preventative and compensatory measures [58]. In many aspects, cyber-attacks bear some similarities with physical faults and transmission errors. DoS attacks can be regarded as coordinated packet dropouts in the sense that both of them impair information availability [64, 83]; the manipulation of sensor measurements can be modeled in a similar way as additive and multiplicative faults [15, 19, 20]. However, existing methods on fault detection and fault-tolerant control can not be applied directly to tackle cyber-attacks. Faults are usually considered as physical events that occur randomly and affect the system performance in an uncoordinated way; whereas cyber-attacks are intentionally designed by attackers, making their detection and mitigation a more challenging task.

Preventative methods are proactive countermeasures that increase the difficulty of launching cyber-attacks or reveal their occurrence at an early stage. Data encryption [67, 72, 95, 96], moving-target defense (MTD) [11, 30, 77], water-marking [53, 61, 63, 78], and novel attack detectors [28, 34, 44, 86] are typical representatives. To resist FDI attacks, the core idea is constructing strict stealthiness conditions such that it is much more expensive or even impossible to launch cyber-attacks without triggering alarms. Data encryption has been widely adopted in computer networks, which relies on the confidentiality of security keys and robustness of encryption algorithms [67]. In data exchange units without enough computing resources, deploying data encryption module could be prohibitive. MTD is a protection technique borrowed from computer security, where a stochastic switching structure is utilized to dynamically and continuously alter the parameters of the system and hinder the attacker's ability to conduct successful reconnaissance. This type of protection requires frequent switching of operation modes and thus is often at the sacrifice of optimal control performance [30]. Water-marking is a physical authentication algorithm that adds a random biased signal to the optimal control input; the manipulation of nominal data can be revealed by checking whether the water-marking signal is correctly reflected in the sensor output [53]. However, one needs to put efforts to balance the attack resilience ability and control performance by designing an optimal authentication signal, since the control input is artificially compromised and the process runs in a non-optimal mode. Designing novel attack detectors without hindering nominal system behavior is a more challenging task. Though there is some work extending fault detection algorithms to reveal cyber-attacks, these methods are effective only for some special types of control systems and data injection [86]. Finally, it is noted that the aforementioned proactive countermeasures are usually not suitable to resist DoS attacks, which are easily notified by system defenders and thus active mitigating methods can be adopted.

As a comparison, compensatory methods refer to those active countermeasures that take effect only after an attack is detected. For industrial control systems, the core mission is to achieve attack-resilient control and estimation, namely, the capacity to maintain system performance at a minimum level even when cyber-attacks are unavoidable. The basic idea is similar to fault tolerant control, where there exist several preset operation modes and robust controllers will be activated after a critical anomaly is announced [94]. It is seen that the attack resilient performance depends on the efficiency of attack detection algorithms. There is also some work assuming that dynamic actions of attackers and defenders are known to each other. As a result, each side will react optimally based on the opponent's optimal action. The decision-making process for both sides is studied in a game-theory framework [17, 35, 37].

1.2 Literature Survey

The thesis focuses on the derivations of worst-case DoS attacks on LQR control channels with energy constraints and FDI attacks on remote state estimation with stealthiness constraints. This section presents a detailed literature survey on the recent development of these attacks and corresponding countermeasures.

1.2.1 Optimal DoS Attacks

In situations where a plant is controlled by a remote controller, both control commands and sensor measurements can be compromised by attackers. Currently, a majority of existing work on DoS attacks concentrates on sensor networks in open-loop systems, where the goal of attackers is to degrade the estimation quality. Zhang *et al.* proved that an optimal DoS strategy against sensor channels of linear quadratic Gaussian (LQG) systems was consecutive attack at active periods when attackers only had limited resource [89]. Li *at al.* formulated a DoS attack and defense strategy in a game-theoretic framework and showed the optimal strategies for sensors and attackers constitute a Nash equilibrium [37]. The authors extended their results to signal-to-interference-plus-noise ratio-based DoS attacks, where a more complex scenario that both the sensor and the attacker can choose their actions with multiple energy levels was investigated [35]. Zhang *at al.* studied the optimal schedule for DoS attacks to degrade the performance of a remote state estimator, where sensor measurements were transmitted through a band-limited wireless channel under the round-robin protocol [90]. For an energy-constraint jammer on remote state estimation, Gan *at al.* investigated the problem of how to select the number of channels at each attack time to maximally deteriorate the CPS performance. Qin *at al.* considered the scenario that an energy-constrained adversary launched DoS attacks on packet-dropping networks and proposed some defensive countermeasures [62]. Compared with attacks on open-loop sensor channels, there are only few papers studying DoS attacks on closed-loop controller links. In general, analysis on the influence of attacks on control performance is more difficult than that on estimation performance because the system dynamics will not be altered in the later case.

Zero-input and hold-input are two compensation strategies under DoS attacks in existing literature [29, 66, 93]. In both strategies, the system dynamics switch between two subsystems and the attack sequence design can be reformulated as an optimal switching problem. Wu *et al.* investigated the FDI optimal attack in LQR systems when attackers had only limited access to actuator channels [82]. The best attacked channels and falsified data were obtained simultaneously by solving a switched LQ problem.

1.2.2 Optimal FDI Attacks

A stealthy FDI attack compromising state estimators was brought forward by Liu *et al.* in smart grids, where it was found that the estimation error of least square estimators could be manipulated arbitrarily large by injecting falsified measurements [45]. Moreover, if the injected data lied in the column space of measurement matrix, the attack could completely deceive a residual-based bad data detector. This interesting conclusion has stimulated substantial investigation on this topic, mainly with the extension from static systems in smart grids to dynamic ones in networked control systems. The capacity of FDI attacks to remain stealthy is the most concerned issue in academic research [22, 51]. In fault detection, the residual (or innovation) is the difference of actual and predicted outputs and often utilized to reveal anomalies [15]. The stealthiness property in existing studies is thus defined in two different ways:

1. From a deterministic point of view, an attack is stealthy if the difference of residuals (or outputs) with and without the attack has a bounded norm [26, 51, 52, 56, 74, 92]. Zero-dynamic deception attacks that can compromise non-minimum phase systems is a typical representative [59].

2. From a statistical point of view, an attack is stealthy if the Kullback–Leibler (KL) divergence of the step-wise innovations with and without the attack is bounded by a given threshold [22, 24, 24, 36, 38–40, 68, 69]. There is also some work utilizing the KL divergence of nominal

and compromised innovation sequences to define a similar stealthiness measure [4, 5, 91].

In both cases, an attack is said to be strictly stealthy if the threshold is set as zero; otherwise the attack is relaxedly stealthy, where more attack performance can be achieved by sacrificing the stealthiness property.

Mo and Sinopoli gave the first residual-based stealthiness definition and proposed the stealthy deception attack for the perfect attackable systems, which could be destabilized, but the residual change was bounded [51]. With the same stealthiness metric, Hu *et al.* found an insecurity condition for the existence of stealthy attacks that could cause unbounded estimation performance degradation [26]. Chen *et al.* investigated the stealthy deception attack with the objective of driving system states to a desired region [10]. A similar problem was studied in [92], where a self-generated stealthy attack leading to unbounded estimation errors was proposed.

The second stealthiness measure is widely applied in optimal deception attacks on remote state estimation, which is motivated by the fact that an interval χ^2 detector is usually deployed to reveal anomalies. Although a lot of research has been carried out on this topic, existing work has been mostly restricted to the case of innovation-based linear attacks. In the pioneering work [22], Guo *et al.* proposed an attack policy that an affine transformation of the nominal innovation was transmitted to the remote end. The optimal coefficients were obtained recursively by solving semi-definite programs (SDPs) that maximize the estimation error subject to stealthiness constraints. If the attacker can intercept only the original measurements, the optimal attack completely deceiving a χ^2 detector was shown to be just flipping the sign of nominal innovations [22]. Guo *et al.* extended their results to the cases where attackers could use extra sensors to measure system states [24] and attacks with relaxed stealthiness [23]. A counter-intuitive conclusion showed that the additional information would not benefit the attacker's purpose for unstable scalar systems. Zhou *et al.* studied a different strategy to utilize side information [97]. All these attacks were linear functions of only the current-step nominal innovation, and thus the compromised innovation was zero-mean independent and identically distributed (i.i.d.). These attacks could successfully deceive an interval χ^2 detector, but they satisfied an overly restrictive stealthiness constraint. To improve attack performance, Li and Yang designed a linear attack based on the current innovation and an additional historical one that was beyond the sliding window of χ^2 detectors [39]. However, the information available to the attacker was not fully utilized, thus the attack did not achieve the maximum estimation quality degradation. To tackle this dilemma, Shang and Chen studied a general scenario that an interval of historical nominal innovations could be utilized to design linear attacks. This policy caused more severe estimation performance loss compared with the previous work, but the compromised innovations were correlated in every two consecutive steps, making the attack only be able to bypass a single-step χ^2 detector [68]. Other relevant papers that studied either the innovation-based linear attacks, or defensive countermeasures against linear attacks can be found in [9, 36, 38, 67, 70, 71, 84]. In summary, it is challenging for innovation-based linear attacks to deceive interval χ^2 detectors without sacrificing attack performance. How to properly handle the stealthiness constraint associated with interval anomaly detectors is still an open problem.

Additionally, very few studies consider the scenario that the attacker can gain some side information of system states by placing extra sensors, though in practical cases this information can be obtained easily. For instance, an adversary may implement another radar to measure the speed of a UAV [18], or deploy an extra thermometer to measure an object's temperature. In [24, 40, 97], different innovation-based linear attacks making use of side information were compared. As will be shown in this thesis, all these policies had an innovation-based linear form and fell short in leveraging the available information efficiently. A systematic understanding of how the additional information affects attack performance is still lacking.

Finally, the above studies focused on the so-call "greedy" attack performance in remote state estimation, i.e., at each step the compromised online measurement is designed only to maximize the current-step estimation error. How to maximize the estimation quality degradation in a fixed interval is a more difficult task. The main challenge is that the compromised innovation has an impact not only on the current-step estimation quality but also on all subsequent estimation errors. Currently, only a few studies considered holistic performance, but all of them were restricted to the case that the attack has an innovation-based linear form [38, 69]. The derivation of optimal attacks maximizing holistic performance without presupposing specific models is still an open problem.

1.2.3 Attack Detection and Resilient Control

It is a challenging task to detect cyber-attacks and take measures to mitigate their impacts, especially for stealthy FDI attacks that can deceive anomaly detectors. In most cases, system security is enhanced with an extra cost, like the tolerable degradation of the optimal control performance. When the widely-used χ^2 detector failed to distinguish the residual under nominal and attacked conditions, Mo *et al.* proposed an active detection method in [53] by intentionally injecting a noise signal into the control input. The so-called watermark signal improved the attack detectability at the cost of increased LQG control effort. Similarly, Romagnoli *et al.* designed a deterministic watermark in [63] using the technique of pseudo-inversion, which could avoid undesirable behavior caused by physical watermarking. Attack resilient estimators could be found in [54, 57], where fundamental problems of reliable state estimation under sensor attacks and bounded noises were investigated. Miao *et al.* proposed a new attack-resilient framework which consisted of multi-combinations of controllers, estimators and detectors for balancing the system's security overhead and control cost [50].

More research work has been done in attack resilient control under DoS attacks. Since system dynamics will be altered in a simple way under DoS, it is natural to formulate the resilient control problem in the framework of switching control. In [31], Lai *et al.* proposed to transmit auxiliary control signals as a single package in every sampling instant. The actuator could use the information contained in the latest received package if it did not receive the command in the next step. By virtue of switched system theory, the proposed method could ensure the closed-loop stability under DoS attacks with maximal length constraints. Kanellopoulos *et al.* established a moving-target defense framework in [30]. The defender designed several optimal controllers with each corresponding to a subset of actuators. The real implemented controller switched among these optimal ones in a random manner. Owing to the moving-target property, it was hard for attackers to identify the effective controller to launch malicious attacks. Recently, Zhu *et al.* studied the observer-based control to stabilize a closed-loop system, assuming the system was subject to periodic DoS attacks in both measurement and control channels [99]. Yong *et al.* modeled the systems under attacks as hidden-model stochastic switching linear systems with unknown inputs, and proposed a multiple-model inference algorithm to tackle security issues [87].

Compared with the synthesis of optimal attacks, the effort in attack detection and resilient control is still lacking. There are few pieces of work studying attack resilient control in the presence of FDI attacks. The "stealthiness" property brings forward the main difficulty in designing effective countermeasures.

1.3 Thesis Contributions

To reveal the vulnerabilities of industrial CPSs, this thesis focuses on the design of optimal cyber-attacks with energy and stealthiness constraints. The major contributions are summarized as follows:

1. We have studied an optimal DoS attack problem against control channels with energy constraints in LQR systems. Two common compensation

strategies under DoS attacks are considered. Necessary and sufficient conditions are derived to ensure attack optimality from initial instants. A general scenario that feasible attacks are not required to be consecutive is also discussed.

2. We have proposed a novel innovation-based attack policy fusing available information and proved that it always outperforms the strategies using only partial information, which clarifies a counter-intuitive conclusion in existing work. More general scenarios are considered, including the correlated measurement noises between two sensors and time-varying means of the injected bias. For attacks that can completely bypass χ^2 detectors, we give explicit solutions for optimal attack policies, which avoid solving optimization problems numerically at each sampling instant.

3. We have revealed that the optimal FDI attack that can compromise Kalman filters and deceive single-step χ^2 detectors is based on the MMSE estimate of prediction errors. A unified design framework and a separation principle are proposed to handle more general attack scenarios, where the attacker may obtain more (or less) measurement data than the remote estimator. The results are extended to the case that multi-step χ^2 detectors are deployed to reveal anomalies.

4. In addition to greedy attack performance, we have also investigated optimal FDI attacks that maximize estimation errors of Kalman filters in a fixed interval. Such information-based attack policies are shown to be a linear function of MMSE estimates of historical prediction errors. The framework covers various scenarios that attackers have access to different levels of online measurements. The optimal stealthy attack compromising a Kalman filter is determined by both the amount of online information and widths of χ^2 detectors.

1.4 Thesis Outline

The remainder of the thesis is organized as follows.

In Chapter 2, the problem of optimal consecutive DoS attacks against LQR control channels with energy constraints is investigated. In Chapter 3, an innovation-based linear FDI attack that compromises remote state estimators is studied when adversaries can gain side information of system states with extra sensors. In Chapter 4, the information-based optimal attack that causes the maximum greedy performance in remote state estimation and deceives a single-step χ^2 detector is proposed. The results are extended to the case that multiple-step χ^2 detectors are deployed to reveal anomalies in Chapter 5. In Chapter 6, the optimal information-based FDI attack leading to the maximum holistic estimation performance loss in Kalman filters is derived. In Chapter 7, concluding remarks of the thesis and some potential directions of future work are provided.

Chapter 2

Optimal DoS Attacks against LQR Control Channels[*]

This chapter investigates the problem of DoS attacks against LQR control channels. Owing to energy constraints, the attacker can only launch consecutive DoS attacks with a certain length m to block communication channels between the controller and actuator. We consider two compensation strategies commonly found in the literature, namely, zero-input and hold-input when control packets are blocked. It is shown that jamming from the initial instant is not always optimal for attackers. Necessary and sufficient conditions are given to ensure the optimality of blocking from the initial instant despite the randomness of initial states. In the case where attackers know the initial state, a finite-interval search method is given to obtain the optimal starting instant of DoS attacks. A general scenario that feasible attacks are not required to be consecutive is also briefly discussed.

This chapter is organized as follows. Section 2.1 formulates the nominal system model and DoS attacks that block control channels. Section 2.2 studies optimal DoS schedules when zero-input and hold-input compensation strategies are adopted by system defenders. Numerical examples are given in Section 2.3 to illustrate the theoretical results. Conclusions are provided in Section 2.4.

[*] A version of this chapter has been published as: Jing Zhou, Jun Shang, Yuzhe Li, and Tongwen Chen, Optimal DoS Attack Against LQR Control Channels. *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 68, no. 4, pp. 1348-1352, April 2021.

2.1 Problem Formulation

In this section, we consider the scenario that an adversary can block the transmission channels of LQR controllers. The nominal and compromised system models are formulated.

2.1.1 Nominal System Model

Consider a discrete-time LTI system given by

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

where $x_k \in \mathbb{R}^n$ is the state, $u_k \in \mathbb{R}^m$ is the control input. Let $Q \succeq 0, R \succ 0$, (A, B) controllable, $(A, Q^{\frac{1}{2}})$ observable. An LQR controller $u_k = Kx_k = -(R + B^{\mathrm{T}}PB)^{-1}B^{\mathrm{T}}PAx_k$ is adopted to minimize the quadratic cost:

$$J = \sum_{k=0}^{\infty}(x_k^{\mathrm{T}}Qx_k + u_k^{\mathrm{T}}Ru_k) \tag{2.2}$$

with P the solution of the Riccati equation:

$$P = A^{\mathrm{T}}PA - A^{\mathrm{T}}PB(R + B^{\mathrm{T}}PB)^{-1}B^{\mathrm{T}}PA + Q. \tag{2.3}$$

The closed-loop system is $x_{k+1} = A_c x_k$ with $A_c = A + BK$. The cost under nominal condition is $J^* = x_0^{\mathrm{T}}Px_0$.

2.1.2 DoS Attack against Control Channels

We assume attackers can obtain system parameters using some techniques, e.g., system identification by intercepting sufficient input/output data, or controller invasion by exploiting system loopholes. Considering Fig. 2.1, attackers can block all communication channels between the controller and actuator in consecutive m steps. The attack performance is measured by (2.2) where u_k is the actual control implemented by actuators. If the attack is not consecutive, the problem is equivalent to designing an optimal switching strategy between two subsystems to maximize the quadratic cost, which has been proven to be

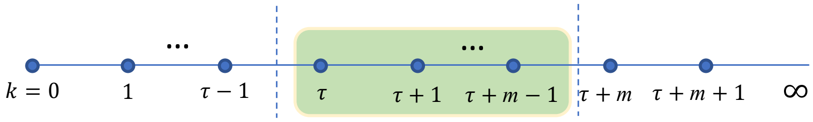

Figure 2.1: Consecutive DoS attack with length m.

The attacker has to decide when to launch the DoS attack to cause the greatest performance loss. Intuitively, an earlier consecutive attack is better than a later one, since the LQR controller will eventually stabilize the system. A too-late attack has little impact on control performance. If an attacker does not know the randomly-configured initial state, he or she should launch DoS attacks from the very beginning. We show this intuitive strategy is not always optimal in the sense of maximizing the performance in (2.2). The optimal delay τ^*, which is determined by system parameters and initial states, is derived based on the available information for attackers.

We explore two simplest compensation strategies commonly found in the literature: the zero-input strategy, where the input to plants is set to zero if a packet is dropped, and the hold-input strategy, where the previous control input is used if a packet is lost [66]. When the control channel is blocked at kth instant, set $u_k = 0$ in Case I and $u_k = u_{k-1}$ in Case II.

2.2 Main Results

As indicated in Fig. 2.1, the time horizon is divided into three subintervals. To calculate the attack performance in (2.2) under DoS attacks, define the mapping $h_X : \mathbb{S}^n \times \mathbb{N} \to \mathbb{S}^n$ as

$$h_X(\Phi, k) = \Phi - (X^k)^{\mathrm{T}} \Phi X^k \tag{2.4}$$

The next two subsections solve the optimal attack problem with zero and hold-input strategies, where the attack performances are denoted by superscript z and h respectively.

2.2.1 Optimal Attack against Zero-Input Strategy

The costs in the first and third intervals are given as

$$J_{\alpha}^{z}(x_0) = x_0^{\mathrm{T}}[P - (A_c^{\tau})^{\mathrm{T}} P A_c^{\tau}]x_0 \tag{2.5}$$

$$J_{\gamma}^{z}(x_0) = x_0^{\mathrm{T}}(A_c^{\tau})^{\mathrm{T}}(A^m)^{\mathrm{T}} P A^m A_c^{\tau} x_0. \tag{2.6}$$

Notice that $u_k = 0$, the cost in the second interval is

$$J_{\beta}^{z}(x_0) = x_0^{\mathrm{T}}(A_c^{\tau})^{\mathrm{T}}[\bar{P} - (A^m)^{\mathrm{T}} \bar{P} A^m] A_c^{\tau} x_0 \tag{2.7}$$

where $\bar{P}$ satisfies the Lyapunov equation

$$\bar{P} = A^{\mathrm{T}} \bar{P} A + Q \tag{2.8}$$

To ensure equation (2.8) has a unique solution, we assume $\bar{\lambda}_i \bar{\lambda}_j \neq 1$, $\forall \bar{\lambda}_i, \bar{\lambda}_j \in \rho_A$, where ρ_A is the spectrum of A. When the open loop system is stable, (2.8) has a positive-definite solution. The cost in the finite interval can be expressed as (2.7). When the open loop system is unstable, (2.7) and (2.8) are still valid. To show this, consider

$$J_{\beta}^{z}(x_0) = x_0^{\mathrm{T}}(A_c^{\tau})^{\mathrm{T}}[\sum_{i=0}^{m-1}(A^i)^{\mathrm{T}} Q A^i] A_c^{\tau} x_0 \tag{2.9}$$

Keep multiplying left and right sides of (2.8) with $(A^i)^{\mathrm{T}}$ and A^i; we have

$$A^{\mathrm{T}} \bar{P} A = (A^2)^{\mathrm{T}} \bar{P} A^2 + A^{\mathrm{T}} Q A$$

$$(A^2)^{\mathrm{T}} \bar{P} A^2 = (A^3)^{\mathrm{T}} \bar{P} A^3 + (A^2)^{\mathrm{T}} Q A^2$$

$$\cdots$$

$$(A^{m-1})^{\mathrm{T}} \bar{P} A^{m-1} = (A^m)^{\mathrm{T}} \bar{P} A^m + (A^{m-1})^{\mathrm{T}} Q A^{m-1}.$$

Summing the left and right sides and canceling identical terms, we have

$$\sum_{i=0}^{m-1}(A^i)^{\mathrm{T}} Q A^i = \bar{P} - (A^m)^{\mathrm{T}} \bar{P} A^m \tag{2.10}$$

which indicates that equations (2.7) and (2.9) are identical.

Define $J_\tau^z(x_0) = J_\alpha^z(x_0) + J_\beta^z(x_0) + J_\gamma^z(x_0)$ as the attack performance when attacks are launched from time τ and the initial state is x_0. In the rest of the brief we denote $J_\tau^z(x_0)$ as J_τ^z for brevity. When attacks start from $\tau = 0$, we have

$$J_0^z = x_0^{\mathrm{T}}[\bar{P} - (A^m)^{\mathrm{T}}\bar{P}A^m + (A^m)^{\mathrm{T}}PA^m]x_0. \tag{2.11}$$

Let $\tilde{P} = \bar{P} - P$, the performance difference between J_0^z and J_τ^z can be obtained by $\Delta J = J_0^z - J_\tau^z = x_0^{\mathrm{T}}W_\tau x_0$, where W_τ is given in the form

$$W_\tau = h_{A_c}[h_A(\tilde{P}, m), \tau] \tag{2.12}$$

Given $m \in \mathbb{N}_+, x_0 \in \mathbb{R}^n$, the attacker needs to find the optimal attack delay τ to maximize J_τ^z, which is equivalent to solving the following optimization problem to minimize ΔJ since J_0^z is a constant:

$$\tau^* = \arg\min_{\tau \in \mathbb{N}} x_0^{\mathrm{T}}W_\tau x_0. \tag{2.13}$$

In reality, the initial state of control systems is randomly configured by system operators and can hardly be obtained by attackers. It is reasonable for attackers to launch an m-step consecutive DoS attack from the initial instant because of the limited information. It is necessary for attackers to find conditions determined by W_τ under which the optimality of blocking from the initial instant can always be preserved despite the randomness of initial states.

Proposition 2.1. *For any $m \in \mathbb{N}_+, h_A(\tilde{P}, m) \succeq 0$.*

Proof. Define $\tilde{Q} = PB(R + B^{\mathrm{T}}PB)^{-1}B^{\mathrm{T}}P$. From (2.3) and (2.8), we have

$$\tilde{P} = A^{\mathrm{T}}\tilde{P}A + A^{\mathrm{T}}\tilde{Q}A. \tag{2.14}$$

Substitute (2.14) into $h_A(\tilde{P}, m)$ and consider the iterative equation

$$h_A(\tilde{P}, m) = \tilde{P} - (A^m)^{\mathrm{T}}\tilde{P}A^m = A^{\mathrm{T}}[h_A(\tilde{P}, m-1) + \tilde{Q}]A. \tag{2.15}$$

Define $\hat{Q} = A^{\mathrm{T}}PA + Q - P = A^{\mathrm{T}}\tilde{Q}A \succeq 0$. Since $h_A(\tilde{P}, 1) = \hat{Q}$, we conclude that for any $m \in \mathbb{N}_+, h_A(\tilde{P}, m) \succeq 0$. ■

When A is stable, (2.14) indicates $\tilde{P} \succ 0$. When A is unstable, $\tilde{P}$ is not positive definite; but Proposition 2.1 shows that $h_A(\tilde{P}, m) \succeq 0$ always holds, which ensures $\Sigma^{-1/2}$ in Theorem 2.1 is well-defined.

Theorem 2.1. *For a given m-step consecutive attack, the attack from initial instant is the best strategy for an arbitrary initial state if and only if $\hat{Q} \succ 0$, $h_A(\tilde{P}, m) = \Psi\Sigma\Psi^{\mathrm{T}}$, $H = \Sigma^{\frac{1}{2}}\Psi^{\mathrm{T}}A_c\Psi\Sigma^{-\frac{1}{2}}$, $\sigma_1(H) \leq 1$.*

Proof. (Sufficiency) Since $\hat{Q} = h_A(\tilde{P}, 1) \succ 0$, by (2.15) we have $h_A(\tilde{P}, m) \succ 0$ for all $m \in \mathbb{N}_+$. Notice that $\Psi\Sigma^{-\frac{1}{2}}\Sigma^{\frac{1}{2}}\Psi^{\mathrm{T}} = I_n$, we have $H^\tau = \Sigma^{\frac{1}{2}}\Psi^{\mathrm{T}}A_c^\tau\Psi\Sigma^{-\frac{1}{2}}$. If $\sigma_1(H) \leq 1$, the following norm inequality holds for any $\tau \in \mathbb{N}_+ \setminus \{1\}$:

$$\sigma_1(H^\tau) \leq \sigma_1(H)\sigma_1(H^{\tau-1}) \leq \sigma_1(H^{\tau-1}) \tag{2.16}$$

which indicates that $\sigma_1(H^\tau)$ is a descending sequence. $\sigma_1(H^\tau) \leq 1$ for any $\tau \in \mathbb{N}_+$. It is equivalent to the condition

$$(H^\tau)^{\mathrm{T}}H^\tau \preceq I_n. \tag{2.17}$$

Considering W_τ in (2.13), we have

$$\begin{aligned} W_\tau = h_{A_c}(\Psi\Sigma\Psi^{\mathrm{T}}, \tau) &= \Psi\Sigma\Psi^{\mathrm{T}} - (A_c^\tau)^{\mathrm{T}}\Psi\Sigma\Psi^{\mathrm{T}}A_c^\tau \\ &= \Psi\Sigma\Psi^{\mathrm{T}} - \Psi\Sigma^{\frac{1}{2}}(H^\tau)^{\mathrm{T}}\Sigma^{-\frac{1}{2}}\Psi^{\mathrm{T}}\Psi\Sigma\Psi^{\mathrm{T}}\Psi\Sigma^{-\frac{1}{2}}H^\tau\Sigma^{\frac{1}{2}}\Psi^{\mathrm{T}} \\ &= \Psi\Sigma^{\frac{1}{2}}[I_n - (H^\tau)^{\mathrm{T}}H^\tau]\Sigma^{\frac{1}{2}}\Psi^{\mathrm{T}} \succeq 0 \end{aligned} \tag{2.18}$$

then for arbitrary $x_0 \in \mathbb{R}^n$, $\Delta J = x_0^{\mathrm{T}}W_\tau x_0 \geq 0$. In this case the optimal attack strategy is to block the control channel from the initial instant.

(Necessity) If the attack from the initial instant is the best choice regardless of x_0, it is necessary to ensure that the attack from $\tau = 0$ is no worse than $\tau = 1$ for any $x_0 \in \mathbb{R}^n$, i.e.,

$$x_0^{\mathrm{T}}W_1x_0 = x_0^{\mathrm{T}}h_{A_c}[h_A(\tilde{P}, m), 1]x_0 \geq 0 \tag{2.19}$$

then we have

$$\begin{aligned} h_{A_c}[h_A(\tilde{P}, m), 1] &= \Psi\Sigma\Psi^{\mathrm{T}} - A_c^{\mathrm{T}}\Psi\Sigma\Psi^{\mathrm{T}}A_c \\ &= \Psi\Sigma^{\frac{1}{2}}(I_n - H^{\mathrm{T}}H)\Sigma^{\frac{1}{2}}\Psi^{\mathrm{T}} \succeq 0. \end{aligned} \tag{2.20}$$

Thus $\sigma_1(H) \leq 1$. ■

Remark 2.1. $\hat{Q} \succ 0$ *is necessary to ensure the optimality of attacking from the initial instant. If* $\hat{Q} \succeq 0$, *let* $x_0 \in Null(\hat{Q})$, *we have* $x_0^{\mathrm{T}} \hat{Q} x_0 = x_0 A^{\mathrm{T}} PB(R+B^{\mathrm{T}} PB)^{-1} B^{\mathrm{T}} PAx_0 = 0$, *then* $Kx_0 = 0$, *indicating the optimal input* $u_0 = 0$. *It is not the best choice to attack from the initial instant since the attacker will waste 'one-step' attack energy. In other words, if* $\hat{Q} \succeq 0$ *the attacker cannot guarantee the strategy to attack from* $\tau = 0$ *is always optimal.*

Notice that $h_{A_c}[h_A(\tilde{P}, m), 0] = 0$, if the condition of Theorem 2.1 is satisfied, the optimal solution of (2.13) is $\tau^* = 0$ for any $x_0 \in \mathbb{R}^n$. The best strategy is blocking control channels from the initial instant in consecutive m steps. The result can be extended to the scenario where the M-step attack is not required to be consecutive, but the best strategy for attackers is still launching an M-step DoS attack consecutively from the beginning. It can be regarded as a special case of the NP-hard problem [85], as concluded in Theorem 2.2.

Theorem 2.2. *For a given total attack step* M, *the* M*-step consecutive attack from the initial instant is the best strategy among all possible attack sequences for an arbitrary initial state if and only if* $\hat{Q} \succ 0$, $h_A(\tilde{P}, m) = \Psi_m \Sigma_m \Psi_m^{\mathrm{T}}$, $H_m = \Sigma_m^{\frac{1}{2}} \Psi_m^{\mathrm{T}} A_c \Psi_m \Sigma_m^{-\frac{1}{2}}$, $\max\{\sigma_1(H_m), m \in [\![1, M]\!]\} \leq 1$.

Proof. (Sufficiency) Suppose an arbitrary attack strategy is randomly choosing M steps in an infinite horizon and setting the control signals to zero. We group the blocked M steps into N subintervals, in each of which the DoS attack is consecutive. Suppose the length of last interval is L_N, $1 \leq L_N \leq M$, the last instant of the penultimate interval is k_s^{N-1}. Since $\max\{\sigma_1(H_m), m \in [\![1, M]\!]\} \leq 1$, we have $\sigma_1(H_{L_N}) \leq 1$. By Theorem 2.1 we know this strategy can be improved by setting the starting point of the last L_N-step consecutive attack as $k_s^{N-1} + 1$, i.e., combining the last two consecutive DoS intervals as a single one yields a better attack strategy. Repeating the reasoning we conclude that the M-step consecutive attack from the initial instant is *always* the optimal strategy for an arbitrary initial state.

(Necessity) Suppose there exists $m \in [\![1, M]\!]$ such that $\sigma_1(H_m) > 1$. It indicates there must be some initial states such that m-step consecutive attack from the initial instant is not optimal. In other words, we can always improve the strategy of consecutive blocking M steps from the initial instant by separating it into two consecutive intervals (with length $M - m$ and m). We conclude that blocking M steps from the initial instant consecutively is not the best attack strategy. ■

Proposition 2.2. *For scalar systems, the M-step consecutive attack from the initial instant is globally optimal for any $M \in \mathbb{N}_+$ and initial states.*

Proof. For a scalar system, for any $m \in [\![1, M]\!]$, $h_A(\tilde{P}, m)$ is a positive real number. Let $h_A(\tilde{P}, m) = \Psi_m \Sigma_m \Psi_m^{\mathrm{T}}$ with $\Psi_m = 1, \Sigma_m = h_A(\tilde{P}, m)$, then $H_m = A_c, \forall m \in [\![1, M]\!]$. Since $\sigma_1(A_c) < 1$, we have $\sigma_1(H_m) < 1$. By Theorem 2.2 we know the optimal attack strategy should be launched from the initial instant for any x_0 and M. ■

Remark 2.2. *Proposition 2.2 indicates that for scalar LQR systems, the attacker should always launch DoS attacks consecutively from the initial instant.*

From the attacker's perspective, if the condition of Theorem 2.1 does not hold, i.e., $\sigma_1(H) > 1$, since W_τ is not positive semidefinite for some τ, there always exists some x_0 such that jamming with τ-step delay yields larger performance loss. The attacker must intercept x_0 to obtain the optimal attack delay. From (2.5)–(2.7), the performance function is rewritten as

$$J_\tau(x_0) = x_0^{\mathrm{T}} P x_0 + x_0^{\mathrm{T}} (A_c^\tau)^{\mathrm{T}} h_A(\tilde{P}, m) A_c^\tau x_0 \qquad (2.21)$$

Let $h_A(\tilde{P}, m) = V^{\mathrm{T}} V$, $J_\tau(x_0) = x_0^{\mathrm{T}} P x_0 + \|V A_c^\tau x_0\|_2^2$. It is not trivial to find $\tau \in \mathbb{N}$ that maximizes $\|V A_c^\tau x_0\|_2$ for a general x_0. The most direct way is brute-force search; but we need to calculate infinite many values since τ can be any positive integer. In the following, we show that only a finite number of calculation is needed to find the optimal τ using the property that the spectral radius of A_c is less than 1. Assume A_c has n independent eigenvectors, denoted as $q_1, ..., q_n$. The corresponding eigenvalues are $\lambda_1, ..., \lambda_n$. For a given $x_0 \in \mathbb{R}^n$ we have $x_0 = \sum_{i=1}^n \mu_i q_i, \mu_i \in \mathbb{R}$. Let $\beta_i = \|\mu_i V q_i\|_2^2$, notice that $|\lambda_i| < 1$, define $\bar{\tau} = \min\{\tau \in \mathbb{N} : \sum_{i=1}^n \beta_i |\lambda_i|^{2\tau} < \sum_{i=1}^n \beta_i\}$. Since $A_c^\tau q_i = \lambda_i^\tau q_i$, when $\tau > \bar{\tau}$, we have

$$\begin{aligned} J_\tau(x_0) &= x_0^{\mathrm{T}} P x_0 + \| \sum_{i=1}^n \lambda_i^\tau \mu_i V q_i \|_2^2 \leq x_0^{\mathrm{T}} P x_0 + \sum_{i=1}^n \beta_i |\lambda_i|^{2\tau} \\ &\leq x_0^{\mathrm{T}} P x_0 + \sum_{i=1}^n \beta_i = J_0(x_0) \end{aligned}$$

which indicates that an m-step consecutive attack starting from $\tau \geq \bar{\tau}$ causes less performance loss than the attack starting from $\tau = 0$. Thus the optimal attack delay can be obtained by

$$\tau^* = \arg \max_{0 \leq \tau \leq \bar{\tau}} \|V A_c^\tau x_0\|_2. \tag{2.22}$$

2.2.2 Optimal Attack against Hold-Input Strategy

In this subsection we consider Case II where the control input will hold constant as the previous step under DoS attacks. The system dynamic in the interval $[\tau, \tau + m - 1]$ is given as

$$\begin{aligned} x_{k+1} &= A x_k + B \bar{u} \\ \bar{u} &= K x_{\tau-1} = K A_c^{\tau-1} x_0 \end{aligned} \tag{2.23}$$

The dynamic can be written as

$$x_{\tau+i} = A^i x_\tau + \sum_{j=0}^{i-1} A^{i-j-1} B K A_c^{\tau-1} x_0 \tag{2.24}$$

Substituting $x_\tau = A_c^\tau x_0$, $A_c = A + BK$ into (24), we have

$$x_{\tau+i} = [A^{i+1} + \sum_{j=0}^{i} (A^j B K)] A_c^{\tau-1} x_0, \quad i \in [\![0, m-1]\!] \tag{2.25}$$

Let $\bar{A}_i = A^{i+1} + \sum_{j=0}^{i} (A^j B K)$, $\Phi_m = \sum_{i=0}^{m-1} (\bar{A}_i^{\mathrm{T}} Q \bar{A}_i) + m K^{\mathrm{T}} R K$. Notice that $x_{\tau+m} = \bar{A}_m A_c^{\tau-1} x_0$, the control performances in the second and third intervals are given by

$$J_\beta^h(x_0) = x_0^{\mathrm{T}} (A_c^{\tau-1})^{\mathrm{T}} \Phi_m A_c^{\tau-1} x_0 \tag{2.26}$$

$$J_\gamma^h(x_0) = x_0^{\mathrm{T}}(A_c^{\tau-1})^{\mathrm{T}}\bar{A}_m^{\mathrm{T}}P\bar{A}_m A_c^{\tau-1}x_0 \tag{2.27}$$

Therefore, the attack performance in Case II is given by

$$\begin{aligned} J_\tau^h = x_0^{\mathrm{T}}[P - (A_c^\tau)^{\mathrm{T}}PA_c^\tau + (A_c^{\tau-1})^{\mathrm{T}}\Phi_m A_c^{\tau-1} \\ + (A_c^{\tau-1})^{\mathrm{T}}\bar{A}_m^{\mathrm{T}}P\bar{A}_m A_c^{\tau-1}]x_0 \end{aligned} \tag{2.28}$$

Attacking from $\tau = 0$ is the same as Case I since $u_{-1} = 0$. Here we consider the special attack launched consecutively from $\tau = 1$; the attack performance becomes

$$J_1^h = x_0^{\mathrm{T}}(P - A_c^{\mathrm{T}}PA_c + \Phi_m + \bar{A}_m^{\mathrm{T}}P\bar{A}_m)x_0 \tag{2.29}$$

Define

$$\Omega_m = \bar{A}_m^{\mathrm{T}}P\bar{A}_m + \Phi_m - A_c^{\mathrm{T}}PA_c \tag{2.30}$$

The performance difference between J_1^h and J_τ^h is given by

$$\Delta J = J_1^h - J_\tau^h = x_0^{\mathrm{T}}h_{A_c}(\Omega_m, \tau - 1)x_0. \tag{2.31}$$

The next proposition shows that Ω_m is also positive semidefinite. Necessary and sufficient conditions to ensure the optimality of attacking from $k = 1$ is given in Theorem 2.3.

Proposition 2.3. *For any $m \in \mathbb{N}_+, \Omega_m \succeq 0$.*

Proof. It is sufficient to show $\Omega_m \succeq \Omega_{m-1}$ and $\Omega_1 \succeq 0$.

$$\begin{aligned} \Omega_m - \Omega_{m-1} &= \bar{A}_m^{\mathrm{T}}P\bar{A}_m - \bar{A}_{m-1}^{\mathrm{T}}P\bar{A}_{m-1} + \Phi_m - \Phi_{m-1} \\ &= \bar{A}_{m-1}^{\mathrm{T}}(A^{\mathrm{T}}PA + Q - P)\bar{A}_{m-1} + K^{\mathrm{T}}B^{\mathrm{T}}PBK \\ &+ \bar{A}_{m-1}^{\mathrm{T}}A^{\mathrm{T}}PBK + K^{\mathrm{T}}B^{\mathrm{T}}PA\bar{A}_{m-1} + K^{\mathrm{T}}RK \end{aligned}$$

Substituting $K = -(R + B^{\mathrm{T}}PB)^{-1}B^{\mathrm{T}}PA$, since $\hat{Q} = A^{\mathrm{T}}PA + Q - P \succeq 0$, we have

$$\begin{aligned} \Omega_m - \Omega_{m-1} &= \bar{A}_{m-1}^{\mathrm{T}}\hat{Q}\bar{A}_{m-1} + \hat{Q} - \bar{A}_{m-1}^{\mathrm{T}}\hat{Q} - \hat{Q}\bar{A}_{m-1} \\ &= (\bar{A}_{m-1} - I)^{\mathrm{T}}\hat{Q}(\bar{A}_{m-1} - I) \succeq 0. \end{aligned} \tag{2.32}$$

Next we consider Ω_1:

$$\begin{aligned}\Omega_1 &= \bar{A}_1^{\mathrm{T}} P \bar{A}_1 + \Phi_1 - A_c^{\mathrm{T}} P A_c \\ &= \bar{A}_1^{\mathrm{T}} P \bar{A}_1 + \bar{A}_0^{\mathrm{T}} Q \bar{A}_0 + K^{\mathrm{T}} R K - A_c^{\mathrm{T}} P A_c \\ &= (A_c - I)^{\mathrm{T}} \hat{Q} (A_c - I) \succeq 0. \end{aligned} \tag{2.33}$$

From (2.32) and (2.33), we have $\Omega_m \succeq 0$ for $m \in \mathbb{N}_+$. ■

Theorem 2.3. *For a given consecutive attack step m, let $\Omega_m = \Psi \Sigma \Psi^{\mathrm{T}}$, $H = \Sigma^{\frac{1}{2}} \Psi^{\mathrm{T}} A_c \Psi \Sigma^{-\frac{1}{2}}$, attack from $\tau = 1$ is the best strategy for arbitrary x_0 if and only if $\hat{Q} \succ 0$, $\sigma_1(H) \leq 1$. For a given total attack step M, let $\Omega_m = \Psi_m \Sigma_m \Psi_m^{\mathrm{T}}$, $H_m = \Sigma_m^{\frac{1}{2}} \Psi_m^{\mathrm{T}} A_c \Psi_m \Sigma_m^{-\frac{1}{2}}$, $m \in [\![1, M]\!]$, the M-step consecutive attack from $\tau = 1$ is the best strategy for an arbitrary initial state if and only if $\hat{Q} \succ 0$, $\max\{\sigma_1(H_m), m \in [\![1, M]\!]\} \leq 1$.*

Proof. Similar to the proofs of Theorems 2.1 and 2.2, and thus omitted. ■

To summarize, the attacker can launch an M-step DoS attack based on the budget. For Case I, if the conditions in Theorem 2.2 hold, the M-step consecutive attack from the initial instant is globally optimal among all feasible attacks for any $x_0 \in \mathbb{R}^n$. If Theorem 2.2 does not hold but the conditions in Theorem 2.1 hold, the M-step consecutive attack from the initial instant is optimal among all consecutive attacks for any $x_0 \in \mathbb{R}^n$. Furthermore, if Theorem 2.1 does not hold, the optimal delay τ^* of M-step consecutive attacks can be obtained by (2.22). Same conclusions can be drawn for Case II, where the optimal strategy is obtained by comparison between $\tau = 0$ and $\tau = 1$ if the conditions in Theorem 2.3 hold.

2.3 Examples

Given $R = I_3$ and system parameters A, B as

$$A = \begin{bmatrix} 1.0305 & -0.0263 \\ 0 & 0.996 \end{bmatrix}, \quad B = \begin{bmatrix} 0.04 & 0.02 & 0.04 \\ 0 & 0.02 & 0.06 \end{bmatrix}$$

Let $Q = \text{diag}(4, q), 0.1 \leq q \leq 0.6$. Consider the zero-hold strategy, the consecutive attack satisfies $1 \leq m \leq 50$. The minimal eigenvalues of W_τ and $\sigma_1(H_m)$ are plotted in Fig. 2.2. As we can see, only when $\sigma_1(H) \leq 1$, the minimal eigenvalue of W_τ is 0, indicating $W_\tau \in \mathbb{S}_+^n$ for any $\tau \in \mathbb{N}$. When $m = 30$, $\sigma_1(H_m) < 1$ for $q \in \{0.4, 0.5, 0.6\}$, then a 30-step consecutive attack from $\tau = 0$ is the best attack strategy. $\sigma_1(H_m) > 1$ for $q \in \{0.1, 0.2, 0.3\}$, let $x_0 = [0.353, 0.936]^{\mathrm{T}}$, the optimal attack delay is obtained by (2.22). The attack performance with different τ is illustrated in Fig. 2.3, where τ^* is marked with a black cross. $J_\tau^z(x_0)$ will converge to J^* as τ increases to infinity, indicating a too-late DoS attack causes negligible impacts on the control performance.

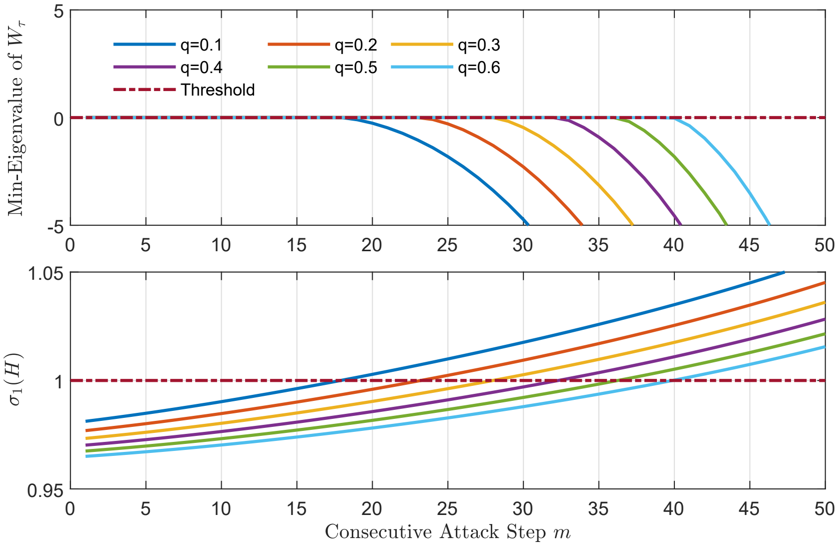

Figure 2.2: Minimal eigenvalue of W_τ and $\sigma_1(H_m)$.

When $q = 0.6, m = 30$, $\sigma_1(H_i) \leq 1\ \forall i \in [\![1, m]\!]$. For this system a 30-step consecutive attack from $\tau = 0$ is globally optimal for all feasible attacks, which are not required to be consecutive.

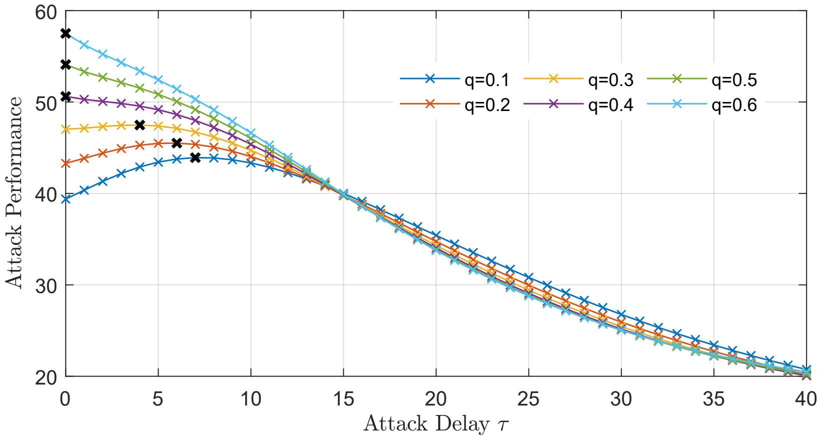

Figure 2.3: 30-step consecutive attack performance.

2.4 Conclusion

This chapter studies the problem of DoS attacks against LQR control channels. Specifically, we demonstrate the counter-intuitive fact that an earlier consecutive DoS attack is not always better than a later one except in scalar LQR systems. The optimal delay of consecutive attacks is given under two compensation strategies. Necessary and sufficient conditions are derived under which the attacker can ensure that consecutive blocking from the initial instant can achieve the greatest performance loss. Future work can be the extension to LQG systems where both controller and sensor channels are vulnerable to DoS attacks.

Chapter 3

Optimal Innovation-Based Linear Deception Attacks with Side Information*

This chapter studies the problem of FDI attacks against remote state estimation. The scenario that malicious attackers can intercept original data packets and also eavesdrop on some side information of system states with extra sensors is considered. To clarify the counter-intuitive issue in existing work, a different innovation-based linear attack policy fusing all available information is proposed. First, the evolution of the *a posteriori* estimation error covariance under FDI attacks is derived. Then, explicit solutions of optimal stealthy attack coefficients are obtained without solving optimization problems numerically. The condition under which there exist multiple optimal attacks is analyzed. Additionally, an easy-to-check criterion for comparing two information fusion methods in scalar systems is given. Simulation results show that, compared with existing work, the proposed attack strategy can completely deceive the anomaly detector and cause more severe performance degradation in remote state estimation.

*A preliminary version of this chapter has been published as: Jing Zhou, Jun Shang, and Tongwen Chen, "Optimal linear FDI attacks with side information: A comparative study," in *4th IEEE International Conference on Industrial Cyber-Physical Systems (ICPS)*, Vancouver, May 2021. A full version entitled "On information fusion in optimal linear FDI attacks against remote state estimation" has been submitted to *IEEE Transactions on Control of Networked Systems* for publication.

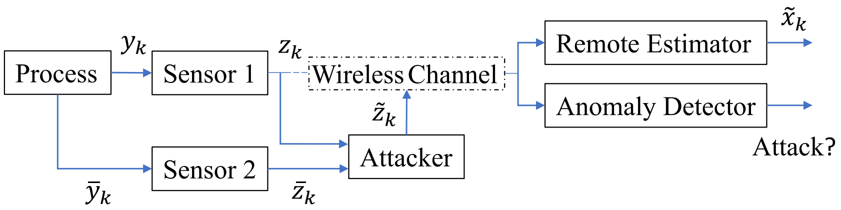

Figure 3.1: Innovation-based FDI attacks with side information.

This chapter is organized as follows. Section 3.1 describes the process model and formulates the optimal attack problem. Section 3.2 studies the estimation performance evolution with FDI attacks. Section 3.3 gives the explicit solutions of optimal attack coefficients. Section 3.4 uses numerical examples and a simplified flight control system to verify the theoretical results. Section 3.5 concludes the chapter.

3.1 Problem Formulation

A discrete-time LTI process is given by

$$x_{k+1} = Ax_k + w_k \tag{3.1}$$

$$y_k = Cx_k + v_k \tag{3.2}$$

where $x_k \in \mathbb{R}^n$ denotes the state vector; $y_k \in \mathbb{R}^m$ is the sensor measurement; $w_k \in \mathbb{R}^n$ and $v_k \in \mathbb{R}^m$ are zero-mean i.i.d. Gaussian noises with covariance $Q \in \mathbb{S}^n_+$ and $R \in \mathbb{S}^m_{++}$, respectively. The initial state x_0 is zero-mean Gaussian with covariance $\Pi_0 \in \mathbb{S}^n_+$, independent of w_k and v_k, $\forall k \in \mathbb{N}$. Assume $m \leq n$ and the pair (A, C) is detectable.

The configuration of remote state estimation is shown in Fig. 3.1. Smart sensor 1 is deployed by system defenders†. At each instant k, it runs a local Kalman filter and sends the innovation $z_k \in \mathbb{R}^m$ to the remote estimator through an unreliable wireless channel [22, 36]. The steady-state Kalman

†Sensor 1 represents the set of all sensors deployed by system defenders but not limited to a single sensor. Same notation applies for Sensor 2.

filter is given as follows:

$$x_{k|k-1} = Ax_{k-1|k-1} \tag{3.3}$$

$$x_{k|k} = x_{k|k-1} + Kz_k \tag{3.4}$$

$$z_k = y_k - Cx_{k|k-1} \tag{3.5}$$

where $x_{k|k-1}$ and $x_{k|k}$ denote the *a priori* and *a posteriori* state estimates, respectively. The optimal state estimate at the remote end is obtained by running a duplicate Kalman filter:

$$\tilde{x}_{k|k-1} = A\tilde{x}_{k-1|k-1} \tag{3.6}$$

$$\tilde{x}_{k|k} = \tilde{x}_{k|k-1} + Kz_k \tag{3.7}$$

with Kalman gain $K = \bar{P}C^{\mathrm{T}}(C\bar{P}C^{\mathrm{T}} + R)^{-1}$ and $\bar{P}$ the solution of Riccati equation $h[g_{[C,R]}(X)] = X$. In the nominal condition, the steady state innovation is zero-mean Gaussian with covariance $\Sigma = C\bar{P}C^{\mathrm{T}} + R$ [22].

3.1.1 Attack Model

We consider a malicious attacker who can obtain all system parameters, intercept the original innovation $\{z_k\}$ and also place an extra sensor (denoted as Sensor 2 in Fig. 3.1) to measure system states:

$$\hat{y}_k = \hat{C}x_k + \hat{v}_k \tag{3.8}$$

with $\hat{C} \in \mathbb{R}^{\bar{m}\times n}$, and $\hat{v}_k$ is a white Gaussian noise with covariance $\hat{R} \in \mathbb{S}^{\bar{m}}_{++}$. Owing to common environmental disturbances, the measurement noises in these two smart sensors may be correlated, i.e., $\mathbb{E}[v_i\hat{v}_j^{\mathrm{T}}] = \delta_{ij}S$ with $S \in \mathbb{R}^{m\times\bar{m}}$. Assume the pair $(A, \hat{C})$ is detectable and the attack starts from $\bar{k}$. The information available to attackers at instant k is denoted by the set $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{z_k, \hat{y}_k\}, \mathbb{I}_{\bar{k}-1} = \emptyset$, where $\{z_k\}$ and $\{\hat{y}_k\}$ refer to intercepted data and side information, respectively. Based on the combined information, the attacker sends fake innovation $\{\tilde{z}_k\}$ to the remote estimator. The attacker's goal is to deteriorate the estimation performance, measured by the *a posteriori* estimation error covariance:

$$\tilde{P}_{k|k} = \mathbb{E}[(x_k - \tilde{x}_{k|k})(x_k - \tilde{x}_{k|k})^{\mathrm{T}}]. \tag{3.9}$$

To obtain the optimal attack sequence, attackers intend to find the mapping $\tilde{z}_k = f(\mathbb{I}_k)$ to maximize $\tilde{P}_k$. In this work, we consider a special attack strategy that $\tilde{z}_k$ is a linear combination of the intercepted innovation and the one based on $\hat{y}_k$, i.e.,

$$\tilde{z}_k = T_k z_k + \hat{T}_k \hat{z}_k + b_k \tag{3.10}$$

where $T_k \in \mathbb{R}^{m\times m}, \hat{T}_k \in \mathbb{R}^{m\times \bar{m}}$; $b_k \in \mathbb{R}^m$ is Gaussian distributed with mean $\mu_k \in \mathbb{R}^m$ and covariance $\Phi_k \in \mathbb{S}_+^m$, $\hat{z}_k \in \mathbb{R}^{\bar{m}}$ is generated by the local Kalman filter of sensor 2:

$$\hat{x}_{k|k-1} = A\hat{x}_{k-1|k-1} \tag{3.11}$$

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + \hat{K}\hat{z}_k \tag{3.12}$$

$$\hat{z}_k = \hat{y}_k - \hat{C}\hat{x}_{k|k-1} \tag{3.13}$$

with the fixed Kalman gain $\hat{K} = \hat{P}\hat{C}^{\mathrm{T}}(\hat{C}\hat{P}\hat{C}^{\mathrm{T}} + \hat{R})^{-1}$ and $\hat{P}$ the solution of $h[g_{[\hat{C},\hat{R}]}(X)] = X$. In steady state, $\hat{z}_k$ is i.i.d. zero-mean Gaussian with covariance $\hat{\Sigma} = \hat{C}\hat{P}\hat{C}^{\mathrm{T}} + \hat{R}$. Let $H_k = \begin{bmatrix} T_k, \hat{T}_k \end{bmatrix} \in \mathbb{R}^{m\times(m+\bar{m})}$ and $\bar{z}_k = \begin{bmatrix} z_k^{\mathrm{T}}, \hat{z}_k^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$; (10) can be rewritten as

$$\tilde{z}_k = H_k \bar{z}_k + b_k,\ b_k \sim \mathcal{N}(\mu_k, \Phi_k). \tag{3.14}$$

3.1.2 Problem of Interest

The attacker cannot launch uncontrolled FDI attacks owing to existence of χ^2 anomaly detectors. To ensure that $\tilde{z}_k$ can successfully bypass the anomaly detector, $\tilde{z}_k$ and z_k must follow a same probability distribution, which imposes the following constraint [22, 24]:

$$\tilde{z}_k \sim \mathcal{N}(0, \Sigma). \tag{3.15}$$

The aim of this work is to derive the optimal stealthy attack strategy, which is given as the solution of the optimization problem:

$$\mathbf{P}_{3.1}: \qquad \max_{H_k, \Phi_k, \mu_k} \quad \mathrm{Tr}(\tilde{P}_{k|k}) \qquad \text{s.t.} \;\; (3.15).$$

Specifically, the following issues should be addressed.

1. How will the estimation performance be degraded under the innovation-based linear attack in (3.14)?

2. What is the optimal policy that can cause the maximum performance loss and also bypass the anomaly detector?

3. Will the additional information for attackers always improve the attack performance?

3.2 Evolution of the Estimation Error Covariance

Attackers can adopt different strategies at each instant to deteriorate the system performance. In this section, we study how the estimation quality will be affected by FDI attacks at the remote end. Assume all Kalman filters have reached steady states before the attack starts. The main result is given in the following theorem.

Theorem 3.1. *The a posteriori estimation error covariance of the remote estimator under innovation-based linear attacks in* (3.14) *evolves according to*

$$\begin{aligned} \tilde{P}_{k|k} = {} & A\tilde{P}_{k-1|k-1}A^{\mathrm{T}} + W + K\tilde{\Sigma}_k K^{\mathrm{T}} + K\mu_k\mu_k^{\mathrm{T}}K^{\mathrm{T}} \\ & -K(H_kY_k - \Omega_k) - (H_kY_k - \Omega_k)^{\mathrm{T}}K^{\mathrm{T}} \end{aligned} \tag{3.16}$$

where

$$\tilde{\Sigma}_k = H_k \Pi H_k^{\mathrm{T}} + \Phi_k, \;\; \Omega_k = \mu_k \sum_{i=\bar{k}}^{k-1} [\mu_i^{\mathrm{T}} K^{\mathrm{T}} (A^{k-i})^{\mathrm{T}}]$$

$$\Pi = \begin{bmatrix} \Sigma & C\Theta^{\mathrm{T}}\hat{C}^{\mathrm{T}} + S \\ \hat{C}\Theta C^{\mathrm{T}} + S^{\mathrm{T}} & \hat{\Sigma} \end{bmatrix}, \;\; Y_k = \begin{bmatrix} CP_k^{\alpha} \\ \hat{C}P_k^{\beta} \end{bmatrix}.$$

The constant matrix $\Theta \in \mathbb{R}^{n\times n}$ is the unique solution of

$$\Theta = (A - A\hat{K}\hat{C})\Theta(A - AKC)^{\mathrm{T}} + A\hat{K}S^{\mathrm{T}}K^{\mathrm{T}}A^{\mathrm{T}} + Q \tag{3.17}$$

and $P_k^{\alpha}, P_k^{\beta}$ are given recursively by

$$\begin{aligned} P_k^{\alpha} &= (A - AKC)(P_{k-1}^{\alpha} - \Theta^{\mathrm{T}}\hat{C}^{\mathrm{T}}\hat{T}_{k-1}^{\mathrm{T}}K^{\mathrm{T}})A^{\mathrm{T}} \\ &\quad + AKS\hat{T}_{k-1}^{\mathrm{T}}K^{\mathrm{T}}A^{\mathrm{T}} + Q \qquad (3.18) \\ P_k^{\beta} &= (A - A\hat{K}\hat{C})(P_{k-1}^{\beta} - \Theta C^{\mathrm{T}}T_{k-1}^{\mathrm{T}}K^{\mathrm{T}})A^{\mathrm{T}} \\ &\quad + A\hat{K}S^{\mathrm{T}}T_{k-1}^{\mathrm{T}}K^{\mathrm{T}}A^{\mathrm{T}} + Q \qquad (3.19) \end{aligned}$$

with initial conditions $P_{\bar{k}-1}^{\alpha} = \bar{P}$, $P_{\bar{k}-1}^{\beta} = \Theta$, $T_{\bar{k}-1} = I_m$, $\hat{T}_{\bar{k}-1} = 0_{m\times\bar{m}}$, $\mu_{\bar{k}-1} = 0_{m\times 1}$, $\Phi_{\bar{k}-1} = 0_m$ and $\tilde{P}_{\bar{k}-1|\bar{k}-1} = (I_n - KC)\bar{P}$.

Proof. Let $e_{k|k-1}$, $\hat{e}_{k|k-1}$ and $\tilde{e}_{k|k-1}$ denote the *a priori* estimation errors of Kalman filters in smart sensor 1, sensor 2 and the remote estimator, respectively. It follows that

$$e_{k+1|k} = (A - AKC)e_{k|k-1} - AKv_k + w_k \tag{3.20}$$

$$\hat{e}_{k+1|k} = (A - A\hat{K}\hat{C})\hat{e}_{k|k-1} - A\hat{K}\hat{v}_k + w_k. \tag{3.21}$$

The remote estimator utilizes the compromised innovation $\tilde{z}_k$ to update state estimation. From (3.10), it can be obtained that

$$\begin{aligned} \tilde{z}_k &= T_k(Ce_{k|k-1} + v_k) + \hat{T}_k(\hat{C}\hat{e}_{k|k-1} + \hat{v}_k) + b_k \\ &= T_kCe_{k|k-1} + \hat{T}_k\hat{C}\hat{e}_{k|k-1} + T_kv_k + \hat{T}_k\hat{v}_k + b_k. \end{aligned} \tag{3.22}$$

From (3.6)–(3.7), we have

$$\begin{aligned} \tilde{e}_{k|k-1} &= x_k - \tilde{x}_{k|k-1} = Ax_{k-1} + w_{k-1} - A\tilde{x}_{k-1|k-1} \\ &= A\tilde{e}_{k-1|k-2} + w_{k-1} - AK\tilde{z}_{k-1}. \end{aligned} \tag{3.23}$$

Substituting (3.22) into (3.23) yields

$$\begin{aligned} \tilde{e}_{k|k-1} &= A\tilde{e}_{k-1|k-2} + w_{k-1} - AK[T_{k-1}(Ce_{k-1|k-2} + v_{k-1}) \\ &\quad + \hat{T}_{k-1}(\hat{C}\hat{e}_{k-1|k-2} + \hat{v}_{k-1}) + b_{k-1}] \\ &= -AKT_{k-1}Ce_{k-1|k-2} - AK\hat{T}_{k-1}\hat{C}\hat{e}_{k-1|k-2} + A\tilde{e}_{k-1|k-2} \\ &\quad - AKT_{k-1}v_{k-1} - AK\hat{T}_{k-1}\hat{v}_{k-1} - AKb_{k-1} + w_{k-1}. \end{aligned} \tag{3.24}$$

Define the state vector $\eta_k \in \mathbb{R}^{3n}$ and matrices

$$\eta_k = \begin{bmatrix} e_{k|k-1} \\ \hat{e}_{k|k-1} \\ \tilde{e}_{k|k-1} \end{bmatrix}, G_k = \begin{bmatrix} A - AKC & 0_n & 0_n \\ 0_n & A - A\hat{K}\hat{C} & 0_n \\ -AKT_kC & -AK\hat{T}_k\hat{C} & A \end{bmatrix}$$

$$F_k = \begin{bmatrix} -AK & 0_{n\times\bar{m}} \\ 0_{m\times n} & -A\hat{K} \\ -AKT_k & -AK\hat{T}_k \end{bmatrix}, M = \begin{bmatrix} I_n \\ I_n \\ I_n \end{bmatrix}, E = \begin{bmatrix} 0_{n\times m} \\ 0_{n\times m} \\ -AK \end{bmatrix}, \bar{R} = \begin{bmatrix} R & S \\ S^{\mathrm{T}} & \hat{R} \end{bmatrix}$$

$$N = \begin{bmatrix} 0_n & 0_n & I_n \end{bmatrix},\ \bar{v}_k = \begin{bmatrix} v_k^{\mathrm{T}} & \hat{v}_k^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}},\ L_k = \begin{bmatrix} T_kC & \hat{T}_k\hat{C} & 0_{m\times n} \end{bmatrix}.$$

Then (3.20)–(3.24) can be written in a compact form

$$\eta_{k+1} = G_k\eta_k + F_k\bar{v}_k + Mw_k + Eb_k \tag{3.25}$$

$$\tilde{z}_k = L_k\eta_k + H_k\bar{v}_k + b_k. \tag{3.26}$$

We now consider the *a posteriori* error covariance of the compromised remote estimator:

$$\begin{aligned} \tilde{P}_{k|k} &= \mathbb{E}[(\tilde{e}_{k|k-1} - K\tilde{z}_k)(\tilde{e}_{k|k-1} - K\tilde{z}_k)^{\mathrm{T}}] \\ &= \mathbb{E}[\tilde{e}_{k|k-1}\tilde{e}_{k|k-1}^{\mathrm{T}}] + K\mathbb{E}[\tilde{z}_k\tilde{z}_k^{\mathrm{T}}]K^{\mathrm{T}} \\ &\quad -K\mathbb{E}[\tilde{z}_k\tilde{e}_{k|k-1}^{\mathrm{T}}] - \mathbb{E}[\tilde{e}_{k|k-1}\tilde{z}_k^{\mathrm{T}}]K^{\mathrm{T}}. \end{aligned} \tag{3.27}$$

Define $P_k^\eta = \mathbb{E}[\eta_k\eta_k^{\mathrm{T}}]$. Note that $\tilde{e}_{k|k-1} = N\eta_k$; the penultimate term of (3.27) satisfies

$$\begin{aligned} \mathbb{E}[\tilde{z}_k\tilde{e}_{k|k-1}^{\mathrm{T}}] &= \mathbb{E}[(L_k\eta_k + H_k\bar{v}_k)(N\eta_k)^{\mathrm{T}} + b_k\tilde{e}_{k|k-1}^{\mathrm{T}}] \\ &= \mathbb{E}[L_k\eta_k\eta_k^{\mathrm{T}}N^{\mathrm{T}} + H_k\bar{v}_k\eta_k^{\mathrm{T}}N^{\mathrm{T}}] + \mathbb{E}[b_k\tilde{e}_{k|k-1}^{\mathrm{T}}] \\ &= L_kP_k^\eta N^{\mathrm{T}} + \mu_k\mathbb{E}[\tilde{e}_{k|k-1}^{\mathrm{T}}]. \end{aligned} \tag{3.28}$$

The last equality is from the facts that $\bar{v}_k$ is zero-mean and independent of η_k, and b_k is independent of $\tilde{e}_{k|k-1}$. Note that P_k^η is a $3n \times 3n$ block matrix, the recursion of which can be derived from (3.25):

$$\begin{aligned} P_{k+1}^\eta = G_kP_k^\eta G_k^{\mathrm{T}} &+ F_k\bar{R}F_k^{\mathrm{T}} + MWM^{\mathrm{T}} + E\mu_k\mu_k^{\mathrm{T}}E^{\mathrm{T}} \\ &+E\Phi_kE^{\mathrm{T}} + G_k\mathbb{E}[\eta_k]\mu_k^{\mathrm{T}}E^{\mathrm{T}} + E\mu_k\mathbb{E}[\eta_k^{\mathrm{T}}]G_k^{\mathrm{T}}. \end{aligned} \tag{3.29}$$

From (3.23), we obtain

$$\mathbb{E}[\tilde{e}_{k|k-1}] = A\mathbb{E}[\tilde{e}_{k-1|k-2}] - AK\mu_{k-1}. \tag{3.30}$$

Because the FDI attack starts from instant $\bar{k}$ and the *a priori* state estimate is not compromised when $k = \bar{k}$, we have $\mathbb{E}[\tilde{e}_{\bar{k}|\bar{k}-1}] = 0_{n\times 1}$. It follows that

$$\mathbb{E}[\tilde{e}_{k|k-1}] = A^{k-\bar{k}}\mathbb{E}[\tilde{e}_{\bar{k}|\bar{k}-1}] - \sum_{i=\bar{k}}^{k-1} A^{k-i}K\mu_i = -\sum_{i=\bar{k}}^{k-1} A^{k-i}K\mu_i. \tag{3.31}$$

The state estimates of two smart sensors are unbiased, leading to $\mathbb{E}[\eta_k] = \left[0_{1\times n},\ 0_{1\times n},\ \mathbb{E}[\tilde{e}^{\mathrm{T}}_{k|k-1}]\right]^{\mathrm{T}}$. The last two terms of (3.29) are obtained from (3.31). Substituting all matrices into (3.29), after some mathematical manipulations, we have

$$\begin{aligned}
P_k^{21} &= (A - A\hat{K}\hat{C})P_{k-1}^{21}(A - AKC)^{\mathrm{T}} \\
&\quad + A\hat{K}S^{\mathrm{T}}K^{\mathrm{T}}A^{\mathrm{T}} + Q && (3.32)\\
P_k^{13} &= (A - AKC)P_{k-1}^{13}A^{\mathrm{T}} + Q - (A - AKC)P_{k-1}^{11}(AKT_{k-1}C)^{\mathrm{T}} \\
&\quad -(A - AKC)P_{k-1}^{12}(AK\hat{T}_{k-1}\hat{C})^{\mathrm{T}} \\
&\quad + AKR(AKT_{k-1})^{\mathrm{T}} + AKS(AK\hat{T}_{k-1})^{\mathrm{T}} && (3.33)
\end{aligned}$$

where P_k^{21} and P_k^{13} are the $(2,1)$ and $(1,3)$ block matrices of P_k^{η}, respectively‡. Since both Kalman filters in two smart sensors are in steady state, P_k^{21} also converges to a constant matrix, given by the unique solution of (3.17). Note that P_k^{11} is the covariance of $e_{k|k-1}$; we have $P_k^{11} = \bar{P}$, then the two terms in P_k^{13} vanish because of the equality:

$$\begin{aligned}
&AKR(AKT_{k-1})^{\mathrm{T}} - (A - AKC)P_{k-1}^{11}(AKT_{k-1}C)^{\mathrm{T}} \\
&= A[K(C\bar{P}C^{\mathrm{T}} + R) - \bar{P}C^{\mathrm{T}}](AKT_{k-1})^{\mathrm{T}} = 0_n.
\end{aligned}$$

Let $P_k^{\alpha} = P_k^{13}$ and substitute $P_k^{11} = \bar{P}, P_k^{12} = \Theta^{\mathrm{T}}$; we obtain the recursion of P_k^{α} in (3.18). Let $P_k^{\beta} = P_k^{23}$; the recursion of P_k^{β} can be derived in a similar way. Now with L_k, N and P_k^{η}, we have

$$L_k P_k^{\eta} N^{\mathrm{T}} = T_k C P_k^{\alpha} + \hat{T}_k\hat{C}P_k^{\beta} = H_k Y_k. \tag{3.34}$$

‡Note that $P_k^{\eta} \in \mathbb{R}^{3n\times 3n}$, which can be partitioned as a 3×3 block matrix. The $(2,1)$ block matrix of P_k^{η} denotes $P_k^{\eta}[n+1:2n, 1:n]$. Same notations apply for P_k^{ij}, $i, j \in [\![1,3]\!]$.

The last two terms of $\tilde{P}_{k|k}$ in (3.27) are obtained by substituting (3.31) and (3.34) into (3.28). From (3.26), the second term of $\tilde{P}_{k|k}$ becomes

$$\begin{aligned}\mathbb{E}[\tilde{z}_k\tilde{z}_k^{\mathrm{T}}] =& L_kP_k^{\eta}L_k^{\mathrm{T}} + H_k\bar{R}H_k^{\mathrm{T}} + \mathbb{E}[b_kb_k^{\mathrm{T}}] \\ =& T_k\Sigma T_k^{\mathrm{T}} + \hat{T}_k\hat{\Sigma}\hat{T}_k^{\mathrm{T}} + T_k(C\Theta^{\mathrm{T}}\hat{C}^{\mathrm{T}} + R)\hat{T}_k^{\mathrm{T}} \\ & + \hat{T}_k(\hat{C}\Theta C^{\mathrm{T}} + R^{\mathrm{T}})T_k^{\mathrm{T}} + \Phi_k + \mu_k\mu_k^{\mathrm{T}} \\ =& H_k\Pi H_k^{\mathrm{T}} + \Phi_k + \mu_k\mu_k^{\mathrm{T}} = \tilde{\Sigma}_k + \mu_k\mu_k^{\mathrm{T}}. \end{aligned} \tag{3.35}$$

The cross-terms vanish because $\bar{v}_k$ is independent of all other variables, and b_k is independent of $e_{k|k-1}$ and $\hat{e}_{k|k-1}$. The first term of $\tilde{P}_{k|k}$ is from the equality $\tilde{P}_{k|k-1} = A\tilde{P}_{k-1|k-1}A^{\mathrm{T}} + Q$ because attacks have no impacts on the prediction step of Kalman filtering. Summarizing the above analysis, the recursion of $\tilde{P}_{k|k}$ is derived and given by (3.16).

When the process is under the nominal condition (without FDI attacks), $e_{k|k-1}$ and $\tilde{e}_{k|k-1}$ are identical. Thus the initial value of P_k^{31} is equal to the steady-state value of P_k^{11}; the initial value of P_k^{23} is equal to the steady-state value of P_k^{21}. We obtain $P_{\bar{k}-1}^{\alpha} = \bar{P}$, $P_{\bar{k}-1}^{\beta} = \Theta$ and $\tilde{P}_{\bar{k}-1|\bar{k}-1} = (I_n - KC)\bar{P}$. Since the attack is launched from $\bar{k}$, we have $T_{\bar{k}-1} = I_m$, $\hat{T}_{\bar{k}-1} = 0_{m\times\bar{m}}$, $\mu_{\bar{k}-1} = 0_{m\times 1}$ and $\Phi_{\bar{k}-1} = 0_m$. ■

Theorem 3.1 shows that the evolution of $\tilde{P}_{k|k}$ involves two-layer recursions. The attack strategy at instant k has explicit impacts on the estimation performance at both current and subsequent steps (by recursions of P_k^{α} and P_k^{β}). It brings additional difficulties if attackers intend to optimize the overall performance degradation in a fixed interval [38]. In this work, we consider only the "greedy" attack strategy that maximizes $\tilde{P}_{k|k}$ given $\tilde{P}_{k-1|k-1}$. Before moving to the design of optimal attacks, we briefly discuss some special cases.

1. If there is no FDI attack, we have $T_k = I_m, \hat{T}_k = 0_{m\times\bar{m}}, b_k = 0_{m\times 1},\ \forall k \in \mathbb{N}$, then $\tilde{\Sigma}_k = \Sigma,\ \Omega_k = 0_{m\times n}$. From (3.19) and (3.34) we have $P_k^{\alpha} = \bar{P}$, $H_kY_k = C\bar{P}$. In this case, $\tilde{P}_{k|k} = (I_n - KC)\bar{P}, \forall k \geq \bar{k}$, which corresponds to a nominal Kalman filter.

2. If $\mu_k = 0_{m\times 1}$ and attackers can only intercept the original innovation, we have $\hat{T}_k = 0_{m\times\bar{m}}, P_k^\alpha = \bar{P}, \forall k \in \mathbb{N}$. It follows that $H_k Y_k = T_k C\bar{P},\ \Omega_k = 0_{m\times n}$, which reduces to Case I in [24].

3. If $\mu_k = 0_{m\times 1}$ and attackers can only obtain side information with extra sensors, we have $T_k = 0_{m\times\bar{m}}, \Omega_k = 0_{m\times n},\ \forall k \in \mathbb{N}$. It follows that

   $$P_k^\beta = (A - A\hat{K}\hat{C})P_{k-1}^\beta A^{\mathrm{T}} + Q.$$

   Then $H_k Y_k = \hat{T}_k \hat{C} P_k^\beta$, which reduces to Case II in [24].

From (3.31), it is observed that the state estimate is biased if $\mathbb{E}[b_k] \neq 0$; then $\tilde{P}_{k|k}$ denotes the second moment of $\tilde{e}_{k|k}$, which can still be used to indicate estimation quality with FDI attacks. From (3.26), the compromised innovation is Gaussian under linear deception attacks, i.e.,

$$\tilde{z}_k \sim \mathcal{N}(\mu_k, H_k \Pi H_k^{\mathrm{T}} + \Phi_k). \tag{3.36}$$

In the above analysis, the compromised innovation is generated by an LTV system in (3.25)–(3.26). The coefficient matrices can be determined offline. However, since the system is driven by white Gaussian noises, this model cannot be used to generate $\tilde{z}_k$ in practical applications. Compared with existing studies, the model facilitates the theoretical analysis of attack performance evolution.

3.3 Optimal Stealthy FDI Attacks

In this section, we consider optimal stealthy FDI attacks. To satisfy the strict stealthiness constraint, attackers must ensure $\mathbb{E}[\tilde{z}_k] = 0_{m\times 1}$, which directly leads to $\mu_k = 0_{m\times 1}$ and $\Omega_k = 0_{m\times n}, \forall k \in \mathbb{N}$. According to Theorem 3.1, the recursion of $\tilde{P}_{k|k}$ has the following form:

$$\tilde{P}_{k|k} = A\tilde{P}_{k-1|k-1}A^{\mathrm{T}} + Q + K\Sigma K^{\mathrm{T}} - KH_k Y_k - Y_k^{\mathrm{T}} H_k^{\mathrm{T}} K^{\mathrm{T}}. \tag{3.37}$$

At the kth sampling instant, the first three terms of $\tilde{P}_{k|k}$ are constant; since the last two terms have the same trace, $\mathbf{P}_{3.1}$ reduces to the optimization problem

$$\begin{aligned} \mathbf{P}_{3.2}: \quad & \min_{H_k,\Phi_k} \quad \mathrm{Tr}(H_k Y_k K) && (3.38)\\ & \text{s.t.} \quad H_k \Pi H_k^{\mathrm{T}} + \Phi_k = \Sigma && (3.39)\\ & \qquad \Phi_k \succeq 0. && (3.40) \end{aligned}$$

Remark 3.1. *In* (3.14)*, we do not presuppose that the injected bias has zero mean. Though it is straightforward to verify that* $\mu_k = 0_{m\times 1}, \forall k \in \mathbb{N}$ *with the stealthiness constraint in* (3.15)*, this assumption facilitates the performance analysis and optimal design of relaxed-stealthy FDI attacks measured by KL divergence [23]. The optimal attack can be obtained by formulating optimization problems. In such cases the assumption that* b_k *is Gaussian with non-zero mean is not conservative, because one can prove that the optimal compromised innovation is indeed Gaussian owing to the fact that Gaussian distribution has the maximal entropy among all probability distributions with the same covariance [23]. In this work we focus on only strictly stealthy attacks.*

3.3.1 Optimal Attack Strategy: Information Fusion I

Although the above semidefinite programs can be solved numerically, attackers need to conduct optimization at each sampling instant, which can be time-consuming. In the following theorem we give the explicit solution of $\mathbf{P}_{3.2}$. It reduces the computation burden and also facilitates analysis on the uniqueness of optimal attack policies.

Theorem 3.2. *The optimal stealthy FDI attack in* (3.14) *is given by*

$$H_k^* = -\Sigma^{\frac{1}{2}}(V_k U_k^{\mathrm{T}} - \tilde{V}_k \mathcal{W}_k \tilde{U}_k^{\mathrm{T}})\Pi^{-\frac{1}{2}}$$

$$\Phi_k^* = \Sigma^{\frac{1}{2}} \tilde{V}_k (I_{m-r_k} - \mathcal{W}_k \mathcal{W}_k^{\mathrm{T}}) \tilde{V}_k^{\mathrm{T}} \Sigma^{\frac{1}{2}}, \ \mu_k^* = 0_{m\times 1}$$

where $r_k = \mathrm{rank}(\Pi^{-\frac{1}{2}} Y_k K \Sigma^{\frac{1}{2}})$*;* U_k, V_k *satisfy the compact singular value decomposition (SVD):*

$$\Pi^{-\frac{1}{2}} Y_k K \Sigma^{\frac{1}{2}} = U_k S_k V_k^{\mathrm{T}}$$

$\tilde{U}_k$ and $\tilde{V}_k$ are orthogonal complements of U_k and V_k, respectively; $\mathcal{W}_k$ is a free parameter satisfying $\mathcal{W}_k\mathcal{W}_k^{\mathrm{T}} \preceq I_{m-r_k}$.

Proof. We start by considering the equality constraint in $\mathbf{P}_{3.2}$. Denote $\bar{H}_k = \Sigma^{-\frac{1}{2}}H_k\Pi^{\frac{1}{2}}$. Left- and right-multiplying (3.39) by $\Sigma^{-\frac{1}{2}}$, we have

$$\bar{H}_k\bar{H}_k^{\mathrm{T}} + \Sigma^{-\frac{1}{2}}\Phi_k\Sigma^{-\frac{1}{2}} = I_m. \tag{3.41}$$

Denote $\bar{Y}_k = \Pi^{-\frac{1}{2}}Y_kK\Sigma^{\frac{1}{2}}$; the objective function becomes

$$\mathrm{Tr}(H_kY_kK) = \mathrm{Tr}(\bar{H}_k\Pi^{-\frac{1}{2}}Y_kK\Sigma^{\frac{1}{2}}) = \mathrm{Tr}(\bar{H}_k\bar{Y}_k).$$

Let $\bar{\Phi}_k = \Sigma^{-\frac{1}{2}}\Phi_k\Sigma^{-\frac{1}{2}}$. According to (3.41), $\mathbf{P}_{3.2}$ becomes

$$\min_{\bar{\Phi}_k\in\mathbb{S}_+^m,\bar{H}_k} \quad \mathrm{Tr}(\bar{H}_k\bar{Y}_k) \quad \text{s.t.} \;\; \bar{H}_k\bar{H}_k^{\mathrm{T}} + \bar{\Phi}_k = I_m. \tag{3.42}$$

Note that $U_k \in \mathbb{R}^{(m+\bar{m})\times r_k}, S_k \in \mathbb{S}_{++}^{r_k}, V_k \in \mathbb{R}^{m\times r_k}$. The objective function satisfies

$$\mathrm{Tr}(\bar{H}_k\bar{Y}_k) = \mathrm{Tr}(\bar{H}_kU_kS_kV_k^{\mathrm{T}}) = \mathrm{Tr}(V_k^{\mathrm{T}}\bar{H}_kU_kS_k) = \sum_{i=1}^{r_k}\tilde{H}_k^{[i,i]}S_k^{[i,i]} \tag{3.43}$$

where $\tilde{H}_k = V_k^{\mathrm{T}}\bar{H}_kU_k \in \mathbb{R}^{r_k\times r_k}$ and $X^{[i,j]}$ denotes the (i,j) entry of X. It is easy to verify that $\tilde{H}_k\tilde{H}_k^{\mathrm{T}} \preceq I_{r_k}$, leading to $\tilde{H}_k^{[i,i]} \in [-1,1], \forall i \in [\![1, r_k]\!]$. From (3.43), we have

$$\mathrm{Tr}(\bar{H}_k\bar{Y}_k) \geq -\sum_{i=1}^{r_k}S_k^{[i,i]} = -\mathrm{Tr}(S_k). \tag{3.44}$$

The equality is attained only when $\tilde{H}_k = V_k^{\mathrm{T}}\bar{H}_kU_k = -I_{r_k}$. Solving this equation, we obtain the optimal solution to (3.42):

$$\bar{H}_k^* = -V_kU_k^{\mathrm{T}} + \mathcal{X}_k\tilde{U}_k^{\mathrm{T}} + \tilde{V}_k\mathcal{Y}_k \tag{3.45}$$

where $\mathcal{X}_k \in \mathbb{R}^{m\times(m+\bar{m}-r_k)}$ and $\mathcal{Y}_k \in \mathbb{R}^{(m-r_k)\times(m+\bar{m})}$ are free parameters. To fulfill the stealthiness constraint in (3.42), it follows that

$$(\mathcal{X}_k\tilde{U}_k^{\mathrm{T}} + \tilde{V}_k\mathcal{Y}_k - V_kU_k^{\mathrm{T}})(\mathcal{X}_k\tilde{U}_k^{\mathrm{T}} + \tilde{V}_k\mathcal{Y}_k - V_kU_k^{\mathrm{T}})^{\mathrm{T}} + \bar{\Phi}_k = I_m. \tag{3.46}$$

Left- and right-multiplying (3.46) by V_k^{T} and V_k, respectively, we have

$$V_k^{\mathrm{T}}\mathcal{X}_k\mathcal{X}_k^{\mathrm{T}}V_k + V_k^{\mathrm{T}}\bar{\Phi}_kV_k = 0_{r_k}.$$

It leads to $V_k^{\mathrm{T}}\mathcal{X}_k = 0_{r_k\times(m+\bar{m}-r_k)}, V_k^{\mathrm{T}}\bar{\Phi}_k = 0_{r_k\times m}$. Left- and right-multiplying (3.46) by V_k^{T} and $\tilde{V}_k$, respectively, we obtain

$$(V_k^{\mathrm{T}}\mathcal{X}_k\tilde{U}_k^{\mathrm{T}} - U_k^{\mathrm{T}})(\tilde{U}_k\mathcal{X}_k^{\mathrm{T}}\tilde{V}_k + \mathcal{Y}_k^{\mathrm{T}}) + V_k^{\mathrm{T}}\bar{\Phi}_k\tilde{V}_k = 0_{r_k\times(m-r_k)}.$$

This implies that $\mathcal{Y}_kU_k = 0_{(m-r_k)\times r_k}$. Let $\bar{\mathcal{X}}_k, \bar{\mathcal{Y}}_k$ be free parameters; $\mathcal{X}_k$ and $\mathcal{Y}_k$ can be parameterized as $\mathcal{X}_k = \tilde{V}_k\bar{\mathcal{X}}_k,\ \mathcal{Y}_k = \bar{\mathcal{Y}}_k\tilde{U}_k^{\mathrm{T}}$. Now (3.45) becomes

$$\bar{H}_k^* = -V_kU_k^{\mathrm{T}} + \tilde{V}_k\mathcal{W}_k\tilde{U}_k^{\mathrm{T}} \tag{3.47}$$

where $\mathcal{W}_k = \bar{\mathcal{X}}_k + \bar{\mathcal{Y}}_k \in \mathbb{R}^{(m-r_k)\times(m+\bar{m}-r_k)}$ is an arbitrary matrix. Notice that $\bar{H}_k^*(\bar{H}_k^*)^{\mathrm{T}} \preceq I_m$, we have

$$V_kV_k^{\mathrm{T}} + \tilde{V}_k\mathcal{W}_k\mathcal{W}_k^{\mathrm{T}}\tilde{V}_k^{\mathrm{T}} \preceq I_m. \tag{3.48}$$

It follows that $\mathcal{W}_k\mathcal{W}_k^{\mathrm{T}} \preceq I_{m-r_k}$. $\bar{\Phi}_k^* = I_m - \bar{H}_k^*(\bar{H}_k^*)^{\mathrm{T}}$. The optimal solution to $\mathbf{P}_{3.2}$ is

$$H_k^* = \Sigma^{\frac{1}{2}}\bar{H}_k^*\Pi^{-\frac{1}{2}},\ \Phi_k^* = \Sigma^{\frac{1}{2}}\bar{\Phi}_k^*\Sigma^{\frac{1}{2}}.$$

The optimal coefficients are obtained with (3.47). ■

Note that $[V_k, \tilde{V}_k]$ and $[U_k, \tilde{U}_k]$ are orthogonal matrices. The simplest way to design $\tilde{V}_k$ and $\tilde{U}_k$ is performing full-size SVD: $\bar{Y}_k = \bar{U}_k\bar{S}_k\bar{V}_k^{\mathrm{T}}$; then let $[U_k, \tilde{U}_k] = \bar{U}_k$, $[V_k, \tilde{V}_k] = \bar{V}_k$. The policy for designing $\tilde{z}_k^*$ is given in Algorithm 3.1. T_k^*, $\hat{T}_k^*$ and Φ_k^* are independent of real-time sensor outputs; thus they can be calculated beforehand to reduce the online computation burden. At each instant, the fake innovation $\tilde{z}_k^*$ is sent to the remote estimator. With this strategy, the attacker can cause the maximum increase in the trace of error covariances, meanwhile remain undetected by the anomaly detector.

Remark 3.2. *If Y_kK has full column rank, i.e., $r_k = m$, then $\mathcal{W}_k$ and $\tilde{V}_k$ vanish. We have $H_k^* = -\Sigma V_kU_k^{\mathrm{T}}\Pi^{-\frac{1}{2}}$, $\Phi_k^* = 0_m$. It follows that $b_k = 0_{m\times 1}$.*

The optimal compromised innovation is unique. If rank$(C) < m$*, we have* $r_k < m$*. The freedom to select* $\mathcal{W}_k$ *leads to multiple optimal solutions, all of which have the same attack performance. For simplicity, one can choose* $\mathcal{W}_k = 0_{(m-r_k)\times(m+\bar{m}-r_k)}$*, leading to*

$$H_k^* = -\Sigma^{\frac{1}{2}} V_k U_k^{\mathrm{T}} \Pi^{-\frac{1}{2}},\ \Phi_k^* = \Sigma^{\frac{1}{2}} \tilde{V}_k \tilde{V}_k^{\mathrm{T}} \Sigma^{\frac{1}{2}}$$

or alternatively design $\mathcal{W}_k$ *such that* $\mathcal{W}_k \mathcal{W}_k^{\mathrm{T}} = I_{m-r_k}$*. This solution yields* $\Phi_k^* = 0_m$*, which eliminates the compensation noisy term in* $\tilde{z}_k$*.*

Remark 3.3. *If attackers can only intercept the original measurement, we have* $\bar{m} = 0, \Pi = \Sigma, \bar{Y}_k = \Sigma^{-\frac{1}{2}} C\bar{K} \Sigma^{\frac{1}{2}} = \Sigma^{-\frac{1}{2}} C \bar{P}^2 C^{\mathrm{T}} \Sigma^{-\frac{1}{2}}$*. It follows that* $\bar{Y}_k = V_k S_k V_k^{\mathrm{T}}, \tilde{U}_k = \tilde{V}_k$*; we can choose* $\mathcal{W}_k = -I_{m-r_k}$*. Then*

$$H_k^* = -\Sigma^{\frac{1}{2}} (V_k V_k^{\mathrm{T}} + \tilde{V}_k \tilde{V}_k^{\mathrm{T}}) \Sigma^{-\frac{1}{2}} = -I_m,\ \Phi_k^* = 0_m.$$

The optimal attack is $\tilde{z}_k^* = H_k^* z_k = -z_k$*. We see that if* rank$(C) = m$*, flipping the sign of nominal innovation [22] is the unique optimal attack. If* rank$(C) < m$*,* $\tilde{z}_k^* = -z_k$ *is only one of the optimal attack policies. The freedom to adopt different optimal policies makes it a more challenging task to design effective countermeasures.*

The following corollary shows that the attack based on combined information always outperforms the ones based on only partial information[§], thus clarifying the counter-intuitive conclusion in [24].

Corollary 3.1. *The attack performance of* (3.14) *based on full information is greater than that of the linear attacks based on only partial information.*

Proof. Easy to verify by noticing that $\tilde{z}_k = T_k z_k + b_k$ and $\tilde{z}_k = \hat{T}_k \hat{z}_k + b_k$ are special cases of (3.14). ■

[§] Partial information refers to the measurement data from only Sensor 1 or Sensor 2. In [24], the authors proved that in some cases, the attacker should only use partial information to design the optimal attack policy.

Algorithm 3.1 Optimal Attacks With Information Fusion I

```
Input: Intercepted data $\{z_k\}$ and side information $\{\bar{y}_k\}$
Output: Optimal compromised innovation $\{\tilde{z}_k^*\}$
Calculate $\Theta$ with (3.17).
Initialize $P^{\alpha}_{\bar{k}-1} = \bar{P}, P^{\beta}_{\bar{k}-1} = \Theta, \tilde{P}_{\bar{k}-1|\bar{k}-1} = (I_n - KC)\bar{P}$.
Set $T_{\bar{k}-1} = I_m, \hat{T}_{\bar{k}-1} = 0_{m\times\bar{m}}$.
for $k = \bar{k} : 1 : \infty$ do
    Run the Kalman filter (3.11)–(3.13) to obtain $\bar{z}_k$.
    Update $P^{\alpha}_k, P^{\beta}_k$ with (3.18)–(3.19).
    Do full-size SVD: $\bar{Y}_k = \bar{U}_k\bar{S}_k\bar{V}_k^{\mathrm{T}}$.
    Set $[U_k, \tilde{U}_k] = \bar{U}_k$, $[V_k, \tilde{V}_k] = \bar{V}_k$. Choose $\mathcal{W}_k$.
    Design $H_k^*, \Phi_k^*$ according to Theorem 3.2.
    Generate compensation noise $b_k \sim \mathcal{N}(0_{m\times 1}, \Phi_k^*)$.
    Design $\tilde{z}_k^*$ with (3.14).
    Evaluate attack performance with (3.37).
end for
```

3.3.2 Optimal Attack Strategy: Information Fusion II

In the combined information case in [24], the optimal compromised innovation is based on the globally optimal state estimation, i.e.,

$$\tilde{z}_k = H_k\check{z}_k + b_k,\ b_k \sim \mathcal{N}(0_{m\times 1}, \Phi_k) \tag{3.49}$$

where $\check{z}_k = \check{y}_k - \check{C}\check{x}_{k|k-1}$; $\check{C} = \left[C^{\mathrm{T}}, \hat{C}^{\mathrm{T}}\right]^{\mathrm{T}}$, $\check{y}_k = \left[y_k^{\mathrm{T}}, \hat{y}_k^{\mathrm{T}}\right]^{\mathrm{T}}$; $\check{x}_{k|k-1}$ is the optimal *a priori* state estimation using combined information and $\check{P}$ is the solution of Riccati equation: $h[g_{[\check{C},\bar{R}]}(X)] = X$. The attack performance is evaluated by

$$\tilde{P}_{k|k} = A\tilde{P}_{k-1|k-1}A^{\mathrm{T}} + Q + K\Sigma K^{\mathrm{T}} - \check{P}\check{C}^{\mathrm{T}}H_k^{\mathrm{T}}K^{\mathrm{T}} - KH_k\check{C}\check{P}. \tag{3.50}$$

Lemma 3.1. *[24, Th. 1] Let* $\check{\Pi} = \check{C}\check{P}\check{C}^{\mathrm{T}} + \bar{R}$*; the optimal attack policy in* (3.49) *is given by the solution of the following optimization problem:*

$$\min_{H_k}\quad \mathrm{Tr}(H_k\check{C}\check{P}K) \quad \text{s.t.}\quad H_k\check{\Pi}H_k^{\mathrm{T}} \preceq \Sigma \tag{3.51}$$

with $\Phi_k^* = \Sigma - H_k^*\check{\Pi}(H_k^*)^{\mathrm{T}}$.

Contrary to $\mathbf{P}_{3.2}$, all parameters in (3.51) are constant at each sampling instant, leading to time-invariant coefficients in (3.49). The following theorem gives the explicit optimal solution.

Theorem 3.3. *The optimal stealthy FDI attack in* (3.49) *is given by*

$$H_k^* = -\Sigma^{\frac{1}{2}}(V_1 U_1^{\mathrm{T}} + \tilde{V}_1 \mathcal{W})\check{\Pi}^{-\frac{1}{2}}$$

$$\Phi_k^* = \Sigma^{\frac{1}{2}}\tilde{V}_1(I_{m-r} - \mathcal{W}\mathcal{W}^{\mathrm{T}})\tilde{V}_1^{\mathrm{T}}\Sigma^{\frac{1}{2}}$$

where U_1, V_1 satisfy the compact SVD:

$$\mathcal{M} = \check{\Pi}^{-\frac{1}{2}}\check{C}\check{P}K\Sigma^{\frac{1}{2}} = U_1 S_1 V_1^{\mathrm{T}},\ S_1 \in \mathbb{S}_{++}^r$$

and $\tilde{V}_1$ is the orthogonal complement of V_1; $\mathcal{W}$ is a free parameter satisfying $\mathcal{W}\mathcal{W}^{\mathrm{T}} \preceq I_{m-r}, \mathcal{W}U_1 = 0_{(m-r)\times r}$.

Proof. Define the Lagrange function associated with (3.51):

$$\mathcal{L}(H_k, \nu_k) = \mathrm{Tr}(H_k\check{C}\check{P}K) + \mathrm{Tr}[\nu_k(H_k\check{\Pi}H_k^{\mathrm{T}} - \Sigma)]$$

where the Lagrange multiplier ν_k is symmetric owing to the symmetry of $H_k\check{\Pi}H_k^{\mathrm{T}} - \Sigma$. The stationary point satisfies

$$(\check{C}\check{P}K)^{\mathrm{T}} + 2\nu_k H_k\check{\Pi} = 0_{m\times(m+\bar{m})} \tag{3.52}$$

$$\nu_k(H_k\check{\Pi}H_k^{\mathrm{T}} - \Sigma) = 0_m. \tag{3.53}$$

Since $\check{\Pi}$ is non-singular, from (3.52), we have $2\Sigma^{\frac{1}{2}}\nu_k H_k\check{\Pi}^{\frac{1}{2}} = -\Sigma^{\frac{1}{2}}(\check{C}\check{P}K)^{\mathrm{T}}\check{\Pi}^{-\frac{1}{2}} = -\mathcal{M}^{\mathrm{T}}$. It follows that

$$4\Sigma^{\frac{1}{2}}\nu_k H_k\check{\Pi}H_k^{\mathrm{T}}\nu_k\Sigma^{\frac{1}{2}} = \mathcal{M}^{\mathrm{T}}\mathcal{M}. \tag{3.54}$$

From (3.53), we have $\Sigma^{\frac{1}{2}}\nu_k H_k\check{\Pi}H_k^{\mathrm{T}}\nu_k\Sigma^{\frac{1}{2}} = \Sigma^{\frac{1}{2}}\nu_k\Sigma\nu_k\Sigma^{\frac{1}{2}}$, then (3.54) becomes

$$4\Sigma^{\frac{1}{2}}\nu_k\Sigma\nu_k\Sigma^{\frac{1}{2}} = \mathcal{M}^{\mathrm{T}}\mathcal{M}.$$

It leads to $2\Sigma^{\frac{1}{2}}\nu_k\Sigma^{\frac{1}{2}} = (\mathcal{M}^{\mathrm{T}}\mathcal{M})^{\frac{1}{2}}$. Note that the other equality is dropped because $\nu_k \in \mathbb{S}_+^m$. Denote $\bar{H}_k = \Sigma^{-\frac{1}{2}}H_k\check{\Pi}^{\frac{1}{2}}$; according to (3.52), we have

$$2\Sigma^{\frac{1}{2}}\nu_k\Sigma^{\frac{1}{2}}\bar{H}_k = -\mathcal{M}^{\mathrm{T}} \tag{3.55}$$

which yields $(\mathcal{M}^{\mathrm{T}}\mathcal{M})^{\frac{1}{2}}\bar{H}_k = -\mathcal{M}^{\mathrm{T}}$. Note that

$$[(\mathcal{M}^{\mathrm{T}}\mathcal{M})^{\frac{1}{2}}]^{+} = V_1 S_1^{-1} V_1^{\mathrm{T}}.$$

The matrix equation yields

$$\bar{H}_k^* = -V_1 S_1^{-1} V_1^{\mathrm{T}} \mathcal{M}^{\mathrm{T}} + \tilde{V}_1 \mathcal{W} = -V_1 U_1^{\mathrm{T}} + \tilde{V}_1 \mathcal{W}$$

where $\mathcal{W} \in \mathbb{R}^{(m-r)\times(m+\bar{m})}$ is an arbitrary matrix. To satisfy the constraint in (3.51), it follows that $\bar{H}_k^*(\bar{H}_k^*)^{\mathrm{T}} \preceq I_m$, i.e.,

$$\begin{bmatrix} V_1 & \tilde{V}_1 \end{bmatrix} \begin{bmatrix} I_r & -U_1^{\mathrm{T}}\mathcal{W}^{\mathrm{T}} \\ -\mathcal{W}U_1 & \mathcal{W}\mathcal{W}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} V_1^{\mathrm{T}} \\ \tilde{V}_1^{\mathrm{T}} \end{bmatrix} \preceq I_m.$$

Thus $\mathcal{W}U_1 = 0_{(m-r)\times r}$, $\mathcal{W}\mathcal{W}^{\mathrm{T}} \preceq I_{m-r}$. Then $\bar{H}_k^* = -V_1 U_1^{\mathrm{T}}$, leading to $H_k^* = \Sigma^{\frac{1}{2}} \bar{H}_k^* \check{\Pi}^{-\frac{1}{2}}$. Substituting $\bar{H}_k^*$ yields the optimal attack coefficients. ■

The optimal attack policy is unique when $\mathrm{rank}(\mathcal{M}) = m$, i.e., $\mathrm{rank}(C) = m$. Multiple optimal solutions exist if $\mathrm{rank}(C) < m$; then the attacker can simply choose $\mathcal{W} = 0_{(m-r)\times(m+\bar{m})}$, leading to

$$H_k^* = -\Sigma^{\frac{1}{2}} V_1 U_1^{\mathrm{T}} \check{\Pi}^{-\frac{1}{2}}, \ \Phi_k^* = \Sigma^{\frac{1}{2}}(I_{m-r} - V_1 V_1^{\mathrm{T}})\Sigma^{\frac{1}{2}}$$

or design $\mathcal{W}$ such that $\mathcal{W}\mathcal{W}^{\mathrm{T}} = I_{m-r}$ to eliminate b_k.

The strategy for designing optimal stealthy attacks in (3.49) is summarized in Algorithm 3.2. The resulting time-invariant policy is easier to implement compared with Algorithm 3.1. But there are also some practical concerns:

1. If Sensor 1 is a smart sensor [22], the nominal innovation is transmitted to the remote end; it is easy to intercept z_k but more challenging to obtain y_k; thus stacking all available measurements to perform optimal state estimation is difficult; Algorithm 3.2 is not applicable in this case ($\check{z}_k$ is not available).

2. With numerical examples, we find that in most cases the policy in Algorithm 3.1 is preferable because it can cause more severe performance loss. Specially, if attackers can obtain only side information with Sensor 2,

Algorithm 3.2 Optimal Attacks With Information Fusion II

```
Input: Intercepted data $\{y_k\}$ and side information $\{\bar{y}_k\}$
Output: Optimal compromised innovation $\{\tilde{z}_k^*\}$
Initialize $\tilde{P}_{\bar{k}-1|\bar{k}-1} = (I_n - KC)\bar{P}$.
Do full-size SVD: $\mathcal{M} = \bar{U}_1 \bar{S}_1 \bar{V}_1^{\mathrm{T}}$.
Set $[U_1, \tilde{U}_1] = \bar{U}_1$, $[V_1, \tilde{V}_1] = \bar{V}_1$.
Choose $\mathcal{W}$. Design $H_k^*, \Phi_k^*$ according to Theorem 3.3.
for $k = \bar{k} : 1 : \infty$ do
    Run the Kalman filter to obtain $\check{z}_k$.
    Generate compensation noise $b_k \sim \mathcal{N}(0_{m\times 1}, \Phi_k^*)$.
    Design $\tilde{z}_k^*$ with (3.49).
    Evaluate attack performance with (3.50).
end for
```

Algorithm 3.1 still yields an optimal policy but Algorithm 3.2 only gives a suboptimal one (the optimal policy is time-varying but Theorem 3.3 yields a time-invariant solution).

In general, comparing attack performance of two different information fusion methods for higher-order systems is difficult, especially when considering "greedy" attack policies. Recall that when we formulate $\mathbf{P}_{3.2}$, the first three terms of $\tilde{P}_{k|k}$ are constant. But if different attack policies are adopted, $\tilde{P}_{k-1|k-1}$ is not consistent; then it is hard to verify whether one policy is always better than another one at each sampling instant. Only in some special cases a lower bound of $\tilde{P}_{k|k}$ can be obtained, which enables us to compare performance of stealthy attacks in (3.14) and (3.49) more efficiently.

3.3.3 Performance Analysis for Scalar Systems

For scalar systems with uncorrelated measurement noises in two smart sensors, the following lemma provides an easy-to-check criterion for selecting the preferable information fusion method.

Lemma 3.2. *Assume $n = m = \bar{m} = 1$ and $S = 0$; let $\check{Y} = \check{C}\check{P}$, $Y = \left[CP, \hat{C}\Theta\right]^{\mathrm{T}}$. The optimal attack in* (3.14) *outperforms the one in* (3.49) *if* *and only if* $Y^{\mathrm{T}}\Pi^{-1}Y > \check{Y}^{\mathrm{T}}\check{\Pi}^{-1}\check{Y}$.

Proof. See Appendix A.1. ■

For unstable scalar systems, the above condition always holds. To demonstrate this, define the following matrix

$$\Delta = \left\{ \begin{bmatrix} \frac{1}{C\bar{P}} & 0 \\ 0 & \frac{1}{\hat{C}\Theta} \end{bmatrix} \Pi \begin{bmatrix} \frac{1}{C\bar{P}} & 0 \\ 0 & \frac{1}{\hat{C}\Theta} \end{bmatrix} \right\}^{-1} = \begin{bmatrix} \frac{\Sigma}{\bar{P}^2C^2} & \frac{1}{\bar{P}} \\ \frac{1}{\bar{P}} & \frac{\Sigma}{\Theta^2\hat{C}^2} \end{bmatrix}^{-1}.$$

If $|A| > 1$, with the matrix inversion lemma and the fact that $\Pi \succeq 0, \bar{P} > \check{P}$, it can be shown that

$$\begin{aligned} Y^{\mathrm{T}}\Pi^{-1}Y = \sum_{i=1,j=1}^{2} \Delta^{[i,j]} \geq \frac{\bar{P}^2C^2}{\Sigma} = \frac{(A^2-1)\bar{P}+W}{A^2} \\ > \frac{(A^2-1)\check{P}+W}{A^2} = \check{Y}^{\mathrm{T}}\check{\Pi}^{-1}\check{Y}. \end{aligned}$$

From Lemma 3.2, the optimal deception attack in (3.14) achieves greater estimation performance loss in unstable scalar systems compared with the one in (3.49).

Remark 3.4. *Define* $\alpha_k = \left[\tilde{e}_{k|k-1}, \hat{z}_k^{\mathrm{T}}\right]^{\mathrm{T}}$, $\beta_k = \left[\tilde{e}_{k|k-1}, \check{z}_k^{\mathrm{T}}\right]^{\mathrm{T}}$. *Recall that* $\tilde{e}_{\bar{k}|\bar{k}-1} = e_{\bar{k}|\bar{k}-1}$. *From the proof of Theorem 3.1, we have*

$$\mathbb{E}[\alpha_{\bar{k}}\alpha_{\bar{k}}^{\mathrm{T}}] = \begin{bmatrix} \bar{P} & Y^{\mathrm{T}} \\ Y & \Pi \end{bmatrix}, \ \mathbb{E}[\beta_{\bar{k}}\beta_{\bar{k}}^{\mathrm{T}}] = \begin{bmatrix} \bar{P} & \check{Y}^{\mathrm{T}} \\ \check{Y} & \check{\Pi} \end{bmatrix}.$$

It follows that

$$\mathbb{E}[\tilde{e}_{\bar{k}|\bar{k}-1}|\hat{z}_{\bar{k}}] = Y^{\mathrm{T}}\Pi^{-1}\hat{z}_{\bar{k}}, \ \mathbb{E}[\tilde{e}_{\bar{k}|\bar{k}-1}|\check{z}_{\bar{k}}] = \check{Y}^{\mathrm{T}}\check{\Pi}^{-1}\hat{z}_{\check{k}}.$$

The variances of the above estimation errors are given by

$$\mathcal{C}_{\alpha} = P - Y^{\mathrm{T}}\Pi^{-1}Y, \ \mathcal{C}_{\beta} = P - \check{Y}^{\mathrm{T}}\check{\Pi}^{-1}\check{Y}.$$

This provides an intuitive explanation for Lemma 3.2: the performance of innovation-based linear attacks depends on the estimation quality for $\tilde{e}_{\bar{k}|\bar{k}-1}$. *It also partially explains why using* $\check{z}_k$ *to design* $\tilde{z}_k$ *does not guarantee to yield an optimal attack [24], because the goals of optimal estimation for* x_k *and* $\tilde{e}_{k|k-1}$ *are not alway consistent.*

3.4 Examples

In this section we use numerical examples to illustrate the theoretical results. Consider a stable system with following parameters:

$$A = \begin{bmatrix} 0.7 & 0.4 & 0 \\ 0 & 0.5 & 0.3 \\ 0 & 0 & 0.7 \end{bmatrix}, \ Q = \operatorname{diag}\left\{ \begin{bmatrix} 0.8 \\ 1.2 \\ 0.5 \end{bmatrix} \right\}, \ C = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix},$$

$$R = \operatorname{diag}\left\{ \begin{bmatrix} 2 \\ 1.2 \end{bmatrix} \right\}, \ \hat{C} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \ \hat{R} = \operatorname{diag}\left\{ \begin{bmatrix} 0.8 \\ 0.5 \end{bmatrix} \right\}.$$

Assume $\bar{k} = 6$. The performance of optimal attacks in (3.14) and (3.49) is illustrated in Fig. 3.2. The stealthy attack from Theorem 3.2, i.e., $\tilde{z}_k$ based on suboptimal state estimations from two smart sensors, can cause greater performance loss compared with that based on globally optimal state estimation. Both policies using combined information outperform the ones using only partial information. Fig. 3.3 shows the performance of FDI attacks on an unstable system (see parameters in [24]). The estimation errors will diverge under all optimal attacks, but the one from Theorem 3.2 has the fastest divergent rate. It can also be observed that the attack from Theorem 3.3 based on combined information causes less performance degradation compared with the ones using only partial information. This comparative result verifies the effectiveness of the information fusion method in Theorem 3.2.

We then consider a simplified linear model of a longitudinal flight control system. The state variables are the pitch angle, pitch rate and velocity (see parameters in [39]). Assume the attacker can use an extra sensor to measure the pitch angle, with $\hat{C} = [1, \ 0, \ 0]$, $\hat{V} = 1.5$. The attack starts from $\bar{k} = 31$. The performance of different policies is plotted in Fig. 3.4. The theoretical evolutions of attack performance are given with (3.37) and (3.50); the empirical ones are obtained by simulating process (3.1)–(3.2) with randomly generated noises for 20,000 times and averaging corresponding square errors at each sampling instant. One can observe that the attack from Theorem 3.2 causes more severe performance loss compared with the one from Theorem 3.3.

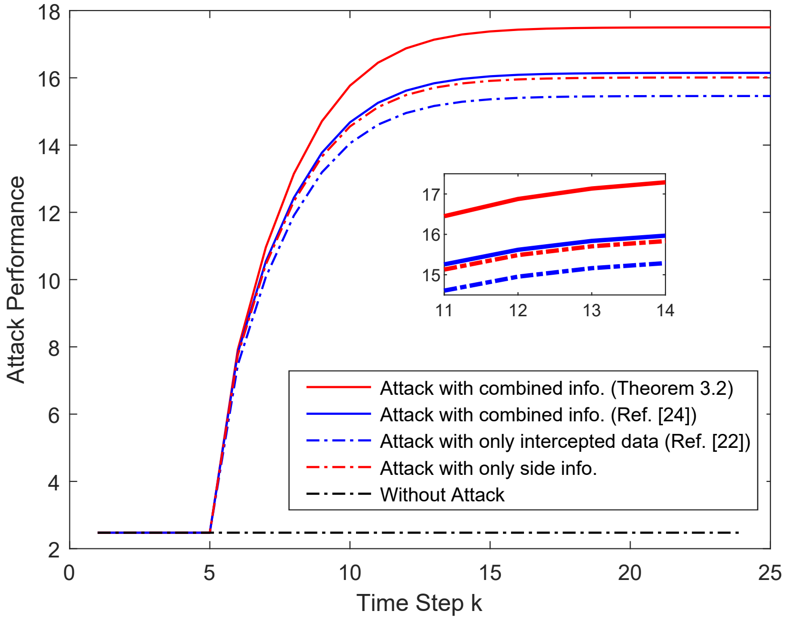

Figure 3.2: Performance of different attack strategies for a stable system.

Fig. 3.5 illustrates the stealthiness property. Set the threshold of the single-step χ^2 detector as 5; the false alarm rate in the nominal condition is 17.18%. The empirical alarm rate with FDI attacks fluctuates in a narrow interval around the theoretical one, indicating that the stealthy attack can completely deceive the anomaly detector.

3.5 Conclusion

In this chapter, we consider the case that a malicious attacker can gain additional information of system states with extra sensors. Using different data fusion methods, the explicit solutions of attack coefficients are derived. The performance of these attack policies are compared using both theoretical justification and simulation examples. Future work will be exploring secure state estimation algorithms to mitigate the impacts of stealthy FDI attacks.

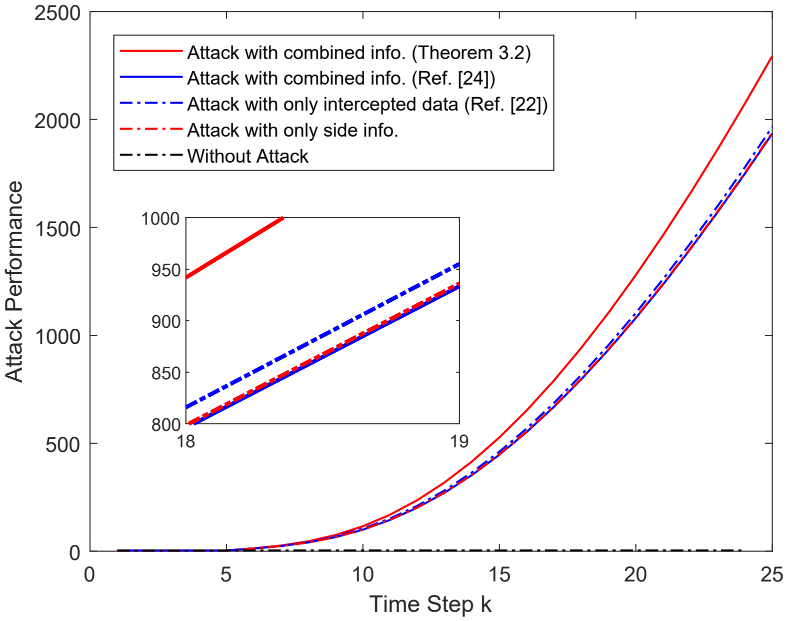

Figure 3.3: Performance of different attack strategies for an unstable system.

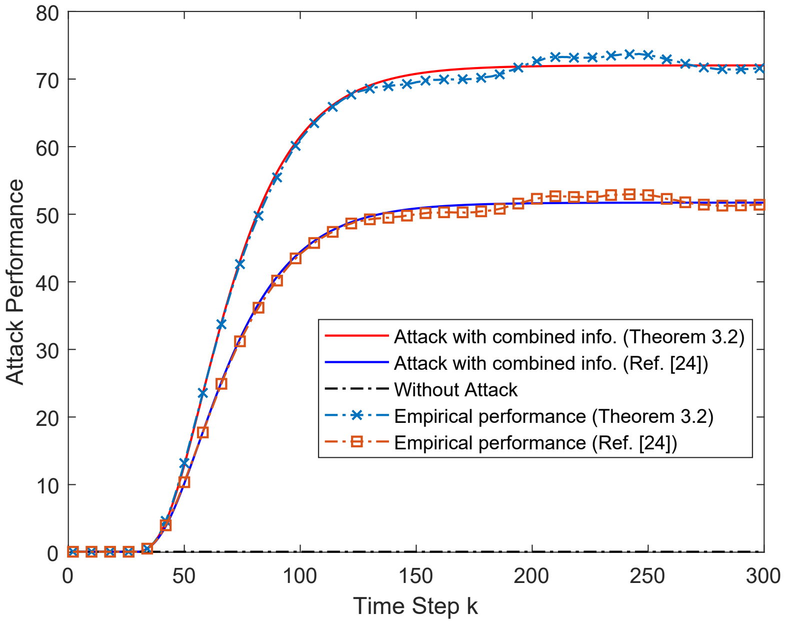

Figure 3.4: Theoretical and empirical attack performance.

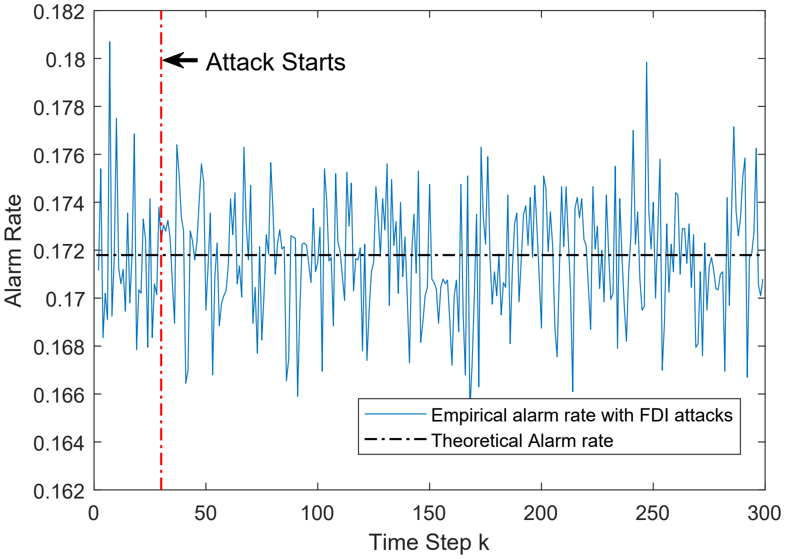

Figure 3.5: Theoretical and empirical alarm rates under FDI attacks.

Chapter 4

Optimal Information-Based Deception Attacks with Single-Step Anomaly Detectors *

This chapter studies the problem of deception attacks against remote state estimation from an information perspective. The Kullback–Leibler divergence between the compromised innovation and nominal one is utilized as the stealthiness measure. Without presupposing a linear attack model, the optimal attack policy that can cause maximum performance loss and deceive the false data detector is derived. For both attacks with strict and relaxed stealthiness, the optimal compromised innovation, which is shown to be generated by a linear time-varying system, can be determined with two steps. First, the minimum mean-square error (MMSE) estimate of the prediction error is obtained using attackers' available information. Then, the faked innovation is designed as a linear transformation of the MMSE estimate. Within a unified framework, this separation principle enables handling more general attack scenarios, where the attacker may obtain more (or less) measurement data than the remote estimator. The optimality of the information-based strategy is verified by theoretical analysis, numerical examples, as well as comparative

*A version of this chapter has been submitted for publication as: Jing Zhou, Jun Shang, and Tongwen Chen, Optimal deception attacks against remote state estimation: An information-based approach, *IEEE Transactions on Automatic Control.* (Accepted as a full paper)

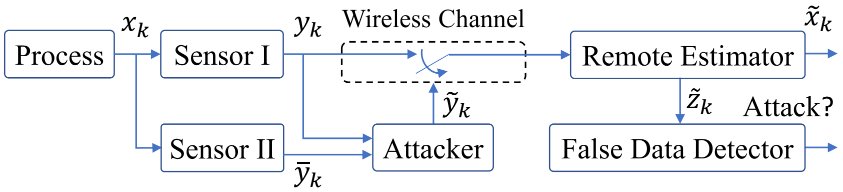

Figure 4.1: Deception attacks against remote state estimation with a single-step χ^2 detector.

studies with existing methods.

This chapter is organized as follows. Section 4.1 describes the system model and formulates the deception attack problem. Section 4.2 focuses on the optimal deception attacks with strict stealthiness. Section 4.3 studies the optimal attacks with relaxed stealthiness. Section 4.4 discusses the scenarios that the attacker has different levels of online information. Section 4.5 provides some numerical examples to verify the theoretical results. Finally, section 4.6 concludes the paper.

4.1 Problem Formulation

The system architecture for remote state estimation is illustrated in Fig. 4.1. The discrete linear time-invariant process is given by (3.1)–(3.2).

4.1.1 Remote Estimator

The measurement y_k is sent sequentially through a wireless channel to the remote end. A standard Kalman filter without time delays and packet loss is deployed to estimate system states [3]:

$$x_{k|k-1} = Ax_{k-1|k-1} \tag{4.1}$$

$$P_{k|k-1} = AP_{k-1|k-1}A^{\mathrm{T}} + Q \tag{4.2}$$

$$K_k = P_{k|k-1}C^{\mathrm{T}} \left(CP_{k|k-1}C^{\mathrm{T}} + R\right)^{-1} \tag{4.3}$$

$$z_k = y_k - Cx_{k|k-1} \tag{4.4}$$

$$x_{k|k} = x_{k|k-1} + K_k z_k \tag{4.5}$$

$$P_{k|k} = (I - K_k C)\, P_{k|k-1} \tag{4.6}$$

where $x_{k|k-1}$ and $x_{k|k}$ denote the *a priori* and *a posteriori* state estimates, respectively; $P_{k|k-1}$ and $P_{k|k}$ are the corresponding estimation error covariances. In steady state, $P_{k|k-1}$ converges to the unique solution of the Riccati equation

$$\bar{P} = h[g_{[C,R]}(\bar{P})]$$

and the nominal innovation $z_k \in \mathbb{R}^m$ is i.i.d. zero-mean Gaussian with covariance $\Sigma = C\bar{P}C^{\mathrm{T}} + R$.

4.1.2 Anomaly Detector

The innovation is sent to the false-data detector at each sampling instant, in order to reveal potential faults or attacks. In this work, we assume that a widely-used χ^2 detector is deployed on the remote side [22–24, 68, 97]. The detector evaluates the following index function:

$$g(z_k) = z_k^{\mathrm{T}} \Sigma^{-1} z_k$$

which is χ^2 distributed with m degrees of freedom. An alarm is raised if $g(z_k)$ exceeds a pre-designed threshold.

4.1.3 Deception Attack

The wireless channel in Fig. 4.1 is unreliable and can be attacked by malicious agents. The attacker can intercept and manipulate y_k. As a result, the compromised measurement $\tilde{y}_k$ is sent to the remote end. Let $\tilde{x}_{k|k-1}$ and

$\tilde{x}_{k|k}$ denote the *a priori* and *a posteriori* remote state estimates with presence of deception attacks, respectively. The corresponding error covariances are defined by

$$\tilde{P}_{k|k-1} = \mathbb{E}[(x_k - \tilde{x}_{k|k-1})(x_k - \tilde{x}_{k|k-1})^{\mathrm{T}}]$$
$$\tilde{P}_{k|k} = \mathbb{E}[(x_k - \tilde{x}_{k|k})(x_k - \tilde{x}_{k|k})^{\mathrm{T}}].$$

The attacker's goal is to cause the maximum deterioration of estimation performance, measured by $\mathrm{Tr}(\tilde{P}_{k|k})$. Assume the attack is launched at the $\bar{k}$th sampling instant; the estimator has entered steady state before $\bar{k}$. To study the worst-case attacks, we make the following assumptions.

Assumption 1. The attacker knows all the system parameters and the state of remote estimator ($\hat{x}_{\bar{k}|\bar{k}-1}$) when attack starts.

Assumption 2. The attacker can eavesdrop on the measurement of Sensor I; they may also obtain some side information of system states by placing extra sensors (denoted by Sensor II in Fig. 4.1). The extra measurement is given by

$$\hat{y}_k = \hat{C}x_k + \hat{v}_k$$

where $\hat{y}_k \in \mathbb{R}^{\bar{m}}, \hat{C} \in \mathbb{R}^{\bar{m}\times n}$; $\hat{v}_k \in \mathbb{R}^{\bar{m}}$ is a zero-mean i.i.d. Gaussian noise with covariance $\hat{R} \in \mathbb{S}^{\bar{m}}_{++}$. The measurement noises of the two sensors may be correlated with covariance $S = \mathbb{E}[v_k \hat{v}_k^{\mathrm{T}}] \in \mathbb{R}^{m\times\bar{m}}$. In practical systems, the side information could be obtained more easily compared with directly intercepting the original measurements. This scenario has not received deserved attention in existing studies. A majority of published papers considered only the case that the attacker can merely modify the measurement of Sensor I; whereas our work studies a more powerful attacker by adding Sensor II.

Remark 4.1. *Assumption 1 is common in the literature on cyber-security [23, 24, 38, 39, 42, 68, 97]. It is in accordance with the Kerckhoffs's principle [73], which stated that the security of a system should not rely on its obscurity. Though it might be difficult in practice to obtain system parameters, we often assume the worst happens that attackers can obtain them using* *techniques like system identification and controller invasion. Stuxnet cyberworm is such a real industrial example [32]. By assuming that attackers have the maximal knowledge of target plants, we can investigate the impact of the worst-case attacks. Additionally, it is not necessary to know the initial state of the estimator in case that smart sensors are used [22]; but if the transmitted data is the raw measurement, the knowledge of $\hat{x}_{\bar{k}|\bar{k}-1}$ is required because the attacker needs an extra Kalman filter to obtain the nominal innovation [23]. Note that $\tilde{x}_{\bar{k}|\bar{k}-1} = \hat{x}_{\bar{k}|\bar{k}-1}$ since the filter is not altered before $\bar{k}$.*

At the kth sampling instant, the information available to the attacker is denoted by the following set:

$$\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k, \hat{y}_k\},\ \forall k \geq \bar{k};\ \mathbb{I}_{\bar{k}-1} = \emptyset. \tag{4.7}$$

Remark 4.2. *With Assumption 2, we can define a general information set. If the attacker cannot eavesdrop on any measurement data, $\mathbb{I}_k = \emptyset$. If the attacker can only intercept the original measurement, $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k\}$. If the attacker can eavesdrop on the original measurement and also obtain some side information by extra sensors, $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k, \hat{y}_k\}$. All these scenarios will be discussed in Section 4.4 as special cases of* (4.7). *Note that a majority of existing work focused only on the second special case.*

4.1.4 Problem of Interest

Owing to the existence of false-data detector, the attacker should design the compromised measurement $\tilde{y}_k$ carefully to remain stealthy. After receiving $\tilde{y}_k$, the innovation becomes

$$\tilde{z}_k = \tilde{y}_k - C\tilde{x}_{k|k-1}. \tag{4.8}$$

Similar to [23], we use the KL divergence between the nominal innovation and compromised one as the stealthiness measure. Let $\delta \in \mathbb{R}_+$ be the stealthiness level defined by attackers; $\tilde{z}_k$ should satisfy the following constraint to deceive the false-data detector:

$$\mathcal{D}_{\mathrm{KL}}(\tilde{z}_k \| z_k) \leq \delta \tag{4.9}$$

where the KL divergence is defined by

$$\mathcal{D}_{\mathrm{KL}}\left(\tilde{z}_k \| z_k\right) = \int_{\mathbb{R}^m} p_{\tilde{z}_k}(t) \ln \frac{p_{\tilde{z}_k}(t)}{p_{z_k}(t)} \mathrm{d}t$$

If the attacker sets $\delta = 0$, then $p_{\tilde{z}_k}(t) = p_{z_k}(t)$, i.e., $\tilde{z}_k$ is also zero-mean Gaussian with covariance Σ. In this case, the attack is strictly stealthy, because the attack detection rate and false alarm rate (FAR) in the nominal condition are the same. If $\delta > 0$, the attack is relaxed-stealthy, in which case more severe performance degradation can be achieved by sacrificing the stealthiness property. It is worth mentioning that evaluating directly the influence of attacks on the alarm rate or $g(z_k)$ is generally a tough task, which may require extensive analysis and computational resources. The cost can be prohibitive for attackers with limited resources. Using the KL divergence as the metric of stealthiness is a common practice [4, 23, 91].

The problem studied in this paper is to derive a stealthy attack sequence to maximize the current-step estimation performance loss, i.e., at each instant, design $\tilde{y}_k$ to maximize $\mathrm{Tr}(\tilde{P}_{k|k})$. This performance criterion is called “greedy” attack performance [22–24, 68]. We formulate it as

$$\mathbf{P}_{4.1}: \quad \max_{\tilde{y}_k = \pi_k(\mathbb{I}_k)} \quad \mathrm{Tr}(\tilde{P}_{k|k}) \quad \text{s.t.} \quad (4.9)$$

where $\pi_k(\mathbb{I}_k)$ denotes the general attack strategy at the kth sampling instant based on all available information. It is known from (4.8) that designing the compromised measurement $\tilde{y}_k$ is equivalent to designing $\tilde{z}_k$ [22, 23]. In the rest of this paper, we use $\tilde{z}_k = \pi_k(\mathbb{I}_k)$ and $\tilde{y}_k = \pi_k(\mathbb{I}_k)$ interchangeably to denote the general attack policy.

Remark 4.3. *In almost all existing work on the same problem, attacks are presupposed to be a linear function of the nominal innovations. By defining $\pi_k(\mathbb{I}_k)$, we do not require that attacks have a specific form. The aim is to find the stealthy attack policy that makes full utilization of available information and achieves the maximum performance degradation.*

4.2 Attacks with Strict Stealthiness

In this section, we study optimal attacks with strict stealthiness. Let $\delta = 0$; the compromised innovation satisfies

$$\tilde{z}_k \sim \mathcal{N}(0, \Sigma). \tag{4.10}$$

4.2.1 MMSE Estimate of the Prediction Error

Denote the *a priori* estimation error (prediction error) with deception attacks as $\tilde{e}_{k|k-1} = x_k - \tilde{x}_{k|k-1}$. The Kalman filter estimates system states recursively by

$$\tilde{x}_{k|k-1} = A\tilde{x}_{k-1|k-1} \tag{4.11}$$

$$\tilde{x}_{k|k} = \tilde{x}_{k|k-1} + K\tilde{z}_k \tag{4.12}$$

where $K = \bar{P}C^{\mathrm{T}}\Sigma^{-1}$ is the steady-state filter gain; $\tilde{z}_k$ is given in (4.8). Substituting (4.11) into (4.12) yields

$$\tilde{x}_{k|k-1} = A\tilde{x}_{k-1|k-1} + AK\tilde{z}_{k-1}. \tag{4.13}$$

Then, we obtain the dynamics of $\tilde{e}_{k|k-1}$ from (3.1) and (4.13):

$$\tilde{e}_{k|k-1} = A\tilde{e}_{k-1|k-1} - AK\tilde{z}_{k-1} + w_{k-1}. \tag{4.14}$$

Define the following matrices:

$$\bar{y}_k = \begin{bmatrix} y_k \\ \hat{y}_k \end{bmatrix}, \bar{v}_k = \begin{bmatrix} v_k \\ \hat{v}_k \end{bmatrix}, \bar{C} = \begin{bmatrix} C \\ \hat{C} \end{bmatrix}, \bar{R} = \begin{bmatrix} R & S \\ S^{\mathrm{T}} & \bar{R} \end{bmatrix}.$$

The measurement data at the kth sampling instant is

$$\bar{y}_k = \bar{C}x_k + \bar{v}_k. \tag{4.15}$$

Define the "virtual sensor" output as

$$r_k = \bar{y}_k - \bar{C}\tilde{x}_{k|k-1}. \tag{4.16}$$

Though there does not exist a real sensor that outputs r_k, the definition helps clarify the subsequent analysis. From (4.13), we have

$$\tilde{x}_{k|k-1} = A^{k-\bar{k}} x_{\bar{k}|\bar{k}-1} + \sum_{i=\bar{k}}^{k-1} A^{k-i} K \tilde{z}_i.$$

This shows that $\tilde{x}_{k|k-1}$ is determined by all historical compromised innovations and $x_{\bar{k}|\bar{k}-1}$. This information is available to the attacker at the kth instant. Since $\bar{y}_k$ is the online measurement, r_k is available to the attacker. Substituting (4.15) into (4.16) yields

$$r_k = \bar{C}\tilde{e}_{k|k-1} + \bar{v}_k. \tag{4.17}$$

Remark 4.4. *The above equation can be regarded as the "virtual measurement" of $\tilde{e}_{k|k-1}$ that is corrupted by a white Gaussian noise with covariance $\bar{R} \in \mathbb{S}_+^{m+\bar{m}}$. The dynamics and measurement of $\tilde{e}_{k|k-1}$ play an important role in deriving the optimal attack policy.*

At the kth sampling instant, $\tilde{z}_{k-1}$ is a known variable. By virtue of (4.14) and (4.17), we use the following Kalman filter to obtain the MMSE estimate of $\tilde{e}_{k|k-1}$:

$$\bar{\xi}_k = A\xi_{k-1} - AK\tilde{z}_{k-1} \tag{4.18}$$

$$\xi_k = \bar{\xi}_k + K_k^{\xi}(r_k - \bar{C}\bar{\xi}_k) \tag{4.19}$$

$$K_k^{\xi} = \bar{P}_k^e \bar{C}^{\mathrm{T}}(\bar{C}\bar{P}_k^e\bar{C}^{\mathrm{T}} + \bar{R})^{-1} \tag{4.20}$$

$$\bar{P}_k^e = AP_{k-1}^e A^{\mathrm{T}} + Q \tag{4.21}$$

$$P_k^e = (I - K_k^{\xi}\bar{C})\bar{P}_k^e \tag{4.22}$$

where $\bar{\xi}_k$ and ξ_k denote the *a priori* and *a posteriori* estimates for $\tilde{e}_{k|k-1}$, respectively. $\bar{P}_k^e$ and P_k^e are the corresponding estimation error covariances, defined by

$$\bar{P}_k^e = \mathbb{E}[(\tilde{e}_{k|k-1} - \bar{\xi}_k)(\tilde{e}_{k|k-1} - \bar{\xi}_k)^{\mathrm{T}}] \tag{4.23}$$

$$P_k^e = \mathbb{E}[(\tilde{e}_{k|k-1} - \xi_k)(\tilde{e}_{k|k-1} - \xi_k)^{\mathrm{T}}]. \tag{4.24}$$

The recursion starts from the $\bar{k}$th sampling instant. Since no measurement information is available before $\bar{k}$, the optimal *a priori* estimate is 0 and the error covariance is $\bar{P}$. Therefore, the initial state of the Kalman filter in (4.18)–(4.22) is $\xi_{\bar{k}} = 0, \bar{P}^e_{\bar{k}} = \bar{P}$. Because the error of MMSE estimation is orthogonal to the estimate, we have

$$\mathbb{E}[\tilde{e}_{k|k-1}\xi_k^{\mathrm{T}}] = \mathbb{E}[\xi_k\xi_k^{\mathrm{T}}] + \mathbb{E}[(\tilde{e}_{k|k-1} - \xi_k)\xi_k^{\mathrm{T}}] = \mathbb{E}[\xi_k\xi_k^{\mathrm{T}}].$$

Define $P^{\xi}_k = \mathbb{E}[\xi_k\xi_k^{\mathrm{T}}]$; expanding (4.24) yields

$$P^{\xi}_k = \tilde{P}_{k|k-1} - P^e_k. \tag{4.25}$$

4.2.2 Optimal Attack Policy

The estimation error covariances evolve according to

$$\tilde{P}_{k|k-1} = A\tilde{P}_{k-1|k-1}A^{\mathrm{T}} + Q \tag{4.26}$$

$$\tilde{P}_{k|k} = \tilde{P}_{k|k-1} + K\mathbb{E}[\tilde{z}_k\tilde{z}_k^{\mathrm{T}}]K^{\mathrm{T}} - K\mathbb{E}[\tilde{z}_k\tilde{e}_{k|k-1}^{\mathrm{T}}] - \mathbb{E}[\tilde{e}_{k|k-1}\tilde{z}_k^{\mathrm{T}}]K^{\mathrm{T}}. \tag{4.27}$$

To evaluate the performance of strictly stealthy attacks, substituting $\mathbb{E}[\tilde{z}_k\tilde{z}_k^{\mathrm{T}}] = \Sigma$ and (4.26) into (4.27), we obtain

$$\tilde{P}_{k|k} = A\tilde{P}_{k-1|k-1}A^{\mathrm{T}} + Q + K\Sigma K^{\mathrm{T}} - K\mathbb{E}[\tilde{z}_k\tilde{e}_{k|k-1}^{\mathrm{T}}] - \mathbb{E}[\tilde{e}_{k|k-1}\tilde{z}_k^{\mathrm{T}}]K^{\mathrm{T}}.$$

At the kth sampling instant, the first three terms of $\tilde{P}_{k|k}$ are constant; hence, maximizing $\mathrm{Tr}(\tilde{P}_{k|k})$ is equivalent to minimizing the trace of the last two terms. We have the reformulated problem as follows:

$$\mathbf{P}_{4.2}: \quad \min_{\tilde{z}_k = \pi_k(\mathbb{I}_k)} \quad \mathrm{Tr}\{K\mathbb{E}[\tilde{z}_k\tilde{e}_{k|k-1}^{\mathrm{T}}]\} \quad \text{s.t.} \quad (4.10).$$

It is worth noting that $\mathbf{P}_{4.2}$ is not a standard convex optimization problem; our purpose is to design the random variable that has a given probability distribution and also minimizes the objective. If it is assumed that $\tilde{z}_k$ is a linear function of z_k, $\mathbf{P}_{4.2}$ can be solved by SDPs [22]. Before we discuss the optimal attack policy, three lemmas are provided.

Lemma 4.1 (see [8, p. 20]). *For $\mathcal{A} \in \mathbb{R}^{n\times n}$, $\mathcal{Q} \in \mathbb{S}^n_{++}$, if*

$$\mathcal{A}^{\mathrm{T}}\mathcal{Q} + \mathcal{Q}\mathcal{A} \preceq 0$$

then $\Re_{\lambda_k}(\mathcal{A}) \leq 0, \forall k \in [\![1, n]\!]$.

Lemma 4.2 (see [7, p. 9]). *For $\mathcal{A} \in \mathbb{S}^n_{+}$, $\mathcal{B} \in \mathbb{S}^n_{+}$, if*

$$\mathcal{A} - \mathcal{B} \succeq 0$$

then $\mathcal{A}^{\frac{1}{2}} - \mathcal{B}^{\frac{1}{2}} \succeq 0$.

Lemma 4.3. *If $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k, \hat{y}_k\}$, or $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k\}$, then $\operatorname{rank}(P^\xi_k) \geq r$, $\operatorname{rank}(K^{\mathrm{T}}P^\xi_k K) = r$.*

Proof. If $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k\}$, we have $\bar{C} = C$, $\bar{R} = R$. The recursions of $\bar{P}^e_k, P^e_k$ and filter gain in (4.18)–(4.22) are the same as the ones in (4.1)–(4.6); then $\bar{P}^e_k$ also converges to $\bar{P}$. Note that $\bar{P}^e_{\bar{k}} = \bar{P}$; (4.18)–(4.22) reduces to a steady-state Kalman filter, leading to

$$K^\xi_k = K,\ P^e_k = (I_n - KC)\bar{P},\ \forall k \geq \bar{k}.$$

If $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k, \hat{y}_k\}$, additional information is used for state estimation. Then $P^e_k \preceq (I_n - KC)\bar{P},\ \forall k \geq \bar{k}$. In the above two cases, we have

$$P^\xi_k = \tilde{P}_{k|k-1} - P^e_k \succeq \tilde{P}_{k|k-1} - \bar{P} + K\Sigma K^{\mathrm{T}}$$
$$K^{\mathrm{T}}P^\xi_k K \succeq K^{\mathrm{T}}(\tilde{P}_{k|k-1} - \bar{P})K + K^{\mathrm{T}}K\Sigma K^{\mathrm{T}}K.$$

Because R is positive definite, Σ is nonsingular; this implies that $\operatorname{rank}(K\Sigma K^{\mathrm{T}}) = \operatorname{rank}(K) = r$. Since deception attacks will not improve the estimation performance, we have $\tilde{P}_{k|k-1} - \bar{P} \in \mathbb{S}^n_{+},\ \forall k \geq \bar{k}$. It follows that $\operatorname{rank}(P^\xi_k) \geq r$, $\operatorname{rank}(K^{\mathrm{T}}P^\xi_k K) = r$. ■

With ξ_k and P^ξ_k obtained from (4.19) and (4.25), the main result in this chapter is given in Theorem 4.1. The case that $\mathbb{I}_k = \emptyset$ will be discussed in Section 4.4.

Theorem 4.1. *If $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k, \hat{y}_k\}$, or $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k\}$, the optimal attack policy with strict stealthiness is given by*

$$\pi_k^*(\mathbb{I}_k): \; \tilde{y}_k^* = C\tilde{x}_{k|k-1} + T_k^* \xi_k + b_k, \; b_k \sim \mathcal{N}(0, \Theta) \tag{4.28}$$

with the coefficients

$$T_k^* = -(V + \bar{V}G)(V^{\mathrm{T}}\Sigma V)^{\frac{1}{2}} V_k U_k^{\mathrm{T}} (U^{\mathrm{T}} P_k^{\xi} U)^{-\frac{1}{2}} U^{\mathrm{T}}$$

$$\Theta = \bar{V}(\bar{V}^{\mathrm{T}}\Sigma\bar{V} - GV^{\mathrm{T}}\Sigma V G^{\mathrm{T}})\bar{V}^{\mathrm{T}}$$

where $[V \; \bar{V}]$ is a unitary matrix; $G = \bar{V}^{\mathrm{T}}\Sigma V(V^{\mathrm{T}}\Sigma V)^{-1}$; U, V and U_k, V_k satisfy the economy-size singular value decompositions (SVD):

$$K = U\hat{S}V^{\mathrm{T}}, \; (U^{\mathrm{T}} P_k^{\xi} U)^{\frac{1}{2}} \hat{S} (V^{\mathrm{T}}\Sigma V)^{\frac{1}{2}} = U_k S_k V_k^{\mathrm{T}}.$$

Proof. The proof is divided into two parts: we first show how to derive the attack policy in (4.28) and then prove its optimality.

Part 1: Note that (4.28) is equivalent to designing the optimal compromised innovation as:

$$\tilde{z}_k^* = T_k^* \xi_k + b_k, \; b_k \sim \mathcal{N}(0, \Theta). \tag{4.29}$$

Define the following variables and matrix:

$$\hat{z}_k = V^{\mathrm{T}}\tilde{z}_k, \; \hat{e}_k = U^{\mathrm{T}}\tilde{e}_{k|k-1}, \; \hat{\Sigma} = V^{\mathrm{T}}\Sigma V \in \mathbb{S}_{++}^r$$

then $\mathbf{P}_{4.2}$ can be reformulated as

$$\min_{\hat{z}_k = \pi_k(\mathbb{I}_k)} \; \mathrm{Tr}\{\hat{S}\mathbb{E}[\hat{z}_k \hat{e}_k^{\mathrm{T}}]\} \quad \text{s.t.} \;\; \hat{z}_k \sim \mathcal{N}(0, \hat{\Sigma}). \tag{4.30}$$

Let $\hat{\xi}_k = U^{\mathrm{T}}\xi_k$ denote the MMSE estimate of $\hat{e}_k$. $\hat{P}_k = U^{\mathrm{T}} P_k^{\xi} U$ is the covariance of $\hat{\xi}_k$. It can be obtained from Lemma 4.3 that $\hat{P}_k \in \mathbb{S}_{++}^r$. We now consider the following attack for (4.30) based on MMSE estimate:

$$\hat{z}_k = \hat{T}_k \hat{\xi}_k + \hat{b}_k, \; \hat{b}_k \sim \mathcal{N}(0, \hat{\Theta}_k) \tag{4.31}$$

where $\hat{b}_k$ is independent of all other variables. $\hat{T}_k$ and $\hat{\Theta}_k$ are adjustable coefficients to ensure stealthiness and optimize the attack performance. Substituting $\hat{z}_k$ into (4.30) yields

$$\begin{aligned} \mathbf{P}_{4.3}: \qquad \min_{\hat{T}_k, \hat{\Theta}_k} \quad & \operatorname{Tr}(\hat{T}_k \hat{P}_k \hat{S}) \\ \text{s.t.} \quad & \hat{T}_k \hat{P}_k \hat{T}_k^{\mathrm{T}} + \hat{\Theta}_k = \hat{\Sigma} \\ & \hat{\Theta}_k \succeq 0. \end{aligned}$$

Denote $\bar{T}_k = \hat{\Sigma}^{-\frac{1}{2}} \hat{T}_k \hat{P}_k^{\frac{1}{2}}, \bar{S}_k = \hat{P}_k^{\frac{1}{2}} \hat{S} \hat{\Sigma}^{\frac{1}{2}}$. Multiply on both sides of the equality constraint by $\hat{\Sigma}^{-\frac{1}{2}}$ and eliminate the slack variable $\hat{\Theta}_k$; then $\mathbf{P}_{4.3}$ becomes

$$\min_{\bar{T}_k} \quad \operatorname{Tr}(\bar{T}_k \bar{S}_k) \quad \text{s.t.} \quad \bar{T}_k \bar{T}_k^{\mathrm{T}} - I_n \preceq 0.$$

Notice that $\bar{S}_k = U_k S_k V_k^{\mathrm{T}}$. The objective function satisfies

$$\operatorname{Tr}(\bar{T}_k \bar{S}_k) = \operatorname{Tr}(V_k^{\mathrm{T}} \bar{T}_k U_k S_k) \geq -\operatorname{Tr}(S_k).$$

Because V_k and U_k are orthogonal matrices and $\bar{T}_k \bar{T}_k^{\mathrm{T}} \preceq I_n$, the minimal value is attained only when $V_k^{\mathrm{T}} \bar{T}_k U_k = -I_n$, i.e., $\bar{T}_k^* = -V_k U_k^{\mathrm{T}}$. For $\mathbf{P}_{4.3}$, we have the unique optimal solution

$$\hat{T}_k^* = -\hat{\Sigma}^{\frac{1}{2}} V_k U_k^{\mathrm{T}} \hat{P}_k^{-\frac{1}{2}}, \ \hat{\Theta}_k^* = 0_r. \tag{4.32}$$

It leads to $\hat{b}_k = 0_r, \forall k \geq \bar{k}$; the optimal linear attack based on MMSE estimate in (6.7) is $\hat{z}_k^* = \hat{T}_k^* \hat{\xi}_k$. Since $\hat{z}_k = V^{\mathrm{T}} \tilde{z}_k$ is not a bijective transformation, one optimal solution to (6.7) corresponds to multiple optimal solutions to $\mathbf{P}_{4.2}$. Solving the matrix equation $\hat{z}_k^* = V^{\mathrm{T}} \tilde{z}_k^*$, the general form of optimal attacks in $\mathbf{P}_{4.2}$ satisfies

$$\tilde{z}_k^* = V \hat{z}_k^* + \bar{V} \epsilon_k \tag{4.33}$$

where $\epsilon_k \in \mathbb{R}^{m-r}$ is an arbitrary random vector. The second term of $\tilde{z}_k^*$ lies in $\operatorname{Ker}(K)$; it has no impact on the attack performance, but will affect the covariance of $\tilde{z}_k^*$. To ensure that (4.33) satisfies the strict stealthiness constraint in $\mathbf{P}_{4.2}$, ϵ_k must be zero-mean Gaussian and the following equality holds:

$$\mathbb{E}[\tilde{z}_k^*(\tilde{z}_k^*)^{\mathrm{T}}] = V\hat{\Sigma}V^{\mathrm{T}} + \bar{V}\mathbb{E}[\epsilon_k\epsilon_k^{\mathrm{T}}]\bar{V}^{\mathrm{T}} + V\mathbb{E}[\hat{z}_k^*\epsilon_k^{\mathrm{T}}]\bar{V}^{\mathrm{T}} \\ +\bar{V}\mathbb{E}[\epsilon_k(\hat{z}_k^*)^{\mathrm{T}}]V^{\mathrm{T}} = \Sigma. \tag{4.34}$$

Now we decompose ϵ_k into two parts, i.e., $\epsilon_k = G\hat{z}_k^* + \bar{\epsilon}_k$. The first part is correlated with $\hat{z}_k^*$ and $\bar{\epsilon}_k \sim \mathcal{N}(0, \Theta_\epsilon)$ is independent of all other variables. Substituting ϵ_k into (4.34) yields

$$(V + \bar{V}G)\hat{\Sigma}(V + \bar{V}G)^{\mathrm{T}} + \bar{V}\Theta_\epsilon\bar{V}^{\mathrm{T}} = \Sigma. \tag{4.35}$$

Note that $\bar{V}^{\mathrm{T}}V = 0_{(m-r)\times r}, V^{\mathrm{T}}V = I_r$ and $\bar{V}^{\mathrm{T}}\bar{V} = I_{m-r}$. Left- and right-multiplying on both sides of (4.35) with $\bar{V}^{\mathrm{T}}$ and V, respectively, we have

$$\bar{V}^{\mathrm{T}}\Sigma V = G\hat{\Sigma}.$$

Multiplying on both sides of (4.35) with $\bar{V}^{\mathrm{T}}$ and $\bar{V}$, we obtain

$$\bar{V}^{\mathrm{T}}\Sigma\bar{V} = G\hat{\Sigma}G^{\mathrm{T}} + \Theta_\epsilon.$$

Since $\hat{\Sigma}$ is nonsingular, G and Θ_ϵ are derived directly from the above two equations. Substituting ϵ_k into (4.33), the compromised innovation becomes

$$\tilde{z}_k^* = (V + \bar{V}G)\hat{z}_k^* + \bar{V}\bar{\epsilon}_k. \tag{4.36}$$

Let $b_k = \bar{V}\bar{\epsilon}_k$; we have $b_k \sim \mathcal{N}(0, \bar{V}\Theta_\epsilon\bar{V}^{\mathrm{T}})$. Substitute $\hat{z}_k^*$ and G into (4.36), then the optimal attack policy $\pi_k^*(\mathbb{I}_k)$ in (4.29) based on MMSE estimate is obtained.

Part 2: Now we prove that $\hat{z}_k^*$ in (4.29) based on the MMSE estimate is optimal among all feasible attacks, which can have an arbitrary form (not necessarily a linear function of z_k). Note that $\pi_k^*(\mathbb{I}_k)$ is a recursive attack policy. At the $\bar{k}$th sampling instant, (4.19) indicates $\xi_{\bar{k}} = K_{\bar{k}}^{\xi}r_{\bar{k}}$; from (4.16) we know $r_{\bar{k}}$ is Gaussian. This implies that $\xi_{\bar{k}}$, and consequently $\tilde{z}_{\bar{k}}^*$ in (4.36), are also Gaussian. By linearity, (4.18)–(4.19) shows that ξ_k is Gaussian distributed if

$\tilde{z}_i^*, \forall i \in [\![\bar{k}, k-1]\!]$ is Gaussian. Therefore, the recursive attack policy satisfies the strict stealthiness constraint.

In (4.33), one can find that $\tilde{z}_k^*$ consists of two parts. $\bar{V}\epsilon_k$ is a compensation term to guarantee stealthiness. Since only $V\hat{z}_k^*$ is effective for performance degradation, it is sufficient to study the attack performance of $\hat{z}_k^*$ for (4.30). Assume that $\hat{z}_k = \pi_k(\mathbb{I}_k)$ satisfying $\hat{z}_k \sim \mathcal{N}(0, \hat{\Sigma})$ is an arbitrary attack strategy. Define the error covariance matrices

$$\hat{P}_k^* = \mathbb{E}[(\hat{z}_k^* - \hat{T}_k^* \hat{e}_k)(\hat{z}_k^* - \hat{T}_k^* \hat{e}_k)^{\mathrm{T}}]$$

$$\hat{P}_k^e = \mathbb{E}[(\hat{z}_k - \hat{T}_k^* \hat{e}_k)(\hat{z}_k - \hat{T}_k^* \hat{e}_k)^{\mathrm{T}}].$$

Note that $\hat{T}_k^*$ is a constant. $\hat{z}_k^*$ is the MMSE estimate for $\hat{T}_k^* \hat{e}_k$. Thus, the following matrix inequality holds[†]:

$$\hat{P}_k^* \preceq \hat{P}_k^e. \tag{4.37}$$

Multiplying on both sides of (4.37) with $\hat{S}$, expanding the inequality and canceling identical terms, we have

$$\begin{aligned} &\hat{S}\hat{T}_k^* \mathbb{E}[\hat{e}_k (\hat{z}_k^*)^{\mathrm{T}}]\hat{S} + \hat{S}\mathbb{E}[\hat{z}^* \hat{e}_k^{\mathrm{T}}](\hat{T}_k^*)^{\mathrm{T}}\hat{S} \\ \succeq &\hat{S}\hat{T}_k^* \mathbb{E}[\hat{e}_k \hat{z}_k^{\mathrm{T}}]\hat{S} + \hat{S}\mathbb{E}[\hat{z}_k \hat{e}_k^{\mathrm{T}}](\hat{T}_k^*)^{\mathrm{T}}\hat{S}. \end{aligned} \tag{4.38}$$

Define $W_k = -\hat{S}\hat{T}_k^*$ and the following matrices associated with the objective function of (4.30):

$$X_k = \mathbb{E}[\hat{e}_k (\hat{z}_k^*)^{\mathrm{T}}]\hat{S}, \ Y_k = \mathbb{E}[\hat{e}_k \hat{z}_k^{\mathrm{T}}]\hat{S}.$$

From $\hat{P}_k^{\frac{1}{2}} \hat{S} \hat{\Sigma}^{\frac{1}{2}} = U_k S_k V_k^{\mathrm{T}}$, we have $\hat{S} = \hat{P}_k^{-\frac{1}{2}} U_k S_k V_k^{\mathrm{T}} \hat{\Sigma}^{-\frac{1}{2}}$. Together with (4.32), it can be obtained that

$$W_k = \hat{P}_k^{-\frac{1}{2}} U_k S_k U_k^{\mathrm{T}} \hat{P}_k^{-\frac{1}{2}} \succ 0. \tag{4.39}$$

From (4.38), we have

$$W_k(X_k - Y_k) + (X_k - Y_k)^{\mathrm{T}} W_k \preceq 0. \tag{4.40}$$

[†] This inequality can be shown by the proof that MMSE estimation is the conditional expectation [3, Th. 3.1].

By Lemma 4.1, (4.39)–(4.40) implies that all eigenvalues of $X_k - Y_k$ have non-positive real parts. Then

$$\mathrm{Tr}(X_k - Y_k) = \sum_{i=1}^{r} \Re_{\lambda_i}(X_k - Y_k) \leq 0.$$

Compared with $\hat{z}_k$, $\hat{z}_k^*$ leads to no larger objective value for (4.30). Since $\hat{z}_k$ is arbitrary by assumption, the above analysis proves the optimality of $\hat{z}_k^*$. ■

If $\mathrm{rank}(C) = m$, K has full column rank; then $\bar{V}$ vanishes. It follows that $\Theta = 0$, $\tilde{z}_k^* = T_k^* \xi_k$, $T_k^* = -V\hat{\Sigma}^{\frac{1}{2}} V_k U_k^{\mathrm{T}} \hat{P}_k^{-\frac{1}{2}} U^{\mathrm{T}}$. The optimal attack policy is unique. If $\mathrm{rank}(C) < m$, there exist multiple optimal policies leading to the same attack performance. Theorem 4.1 gives the simplest one. The first reason is that in (4.33) we can design ϵ_k with a general form, i.e., $\epsilon_k = h_k(\mathbb{I}_k) + \bar{\epsilon}_k$. $h_k(\mathbb{I}_k)$ is the mapping (possibly nonlinear) that reflects the correlation of ϵ_k with $\hat{z}_k^*$ and $\bar{\epsilon}_k$ is an independent term. Any $h_k(\mathbb{I}_k)$ that ensures ϵ_k is Gaussian and satisfies (4.34) corresponds to an optimal solution to $\mathbf{P}_{4.2}$. Since ϵ_k does not affect the attack performance, we can simply design $h_k(\cdot)$ as a linear function of $\hat{z}_k^*$, which gives the result in (4.28). The freedom to design $\bar{V}$ also leads to the non-uniqueness of $\tilde{z}_k^*$. Note that the columns of $\bar{V}$ form an orthogonal basis for $\mathrm{Ker}(K)$. One can do full-size SVD: $K = \tilde{U}\tilde{S}\tilde{V}^{\mathrm{T}}$, then construct V and $\bar{V}$ as the first r and last $m - r$ columns of $\tilde{V}$, respectively.

Remark 4.5. *It is known that MMSE estimation is the conditional expectation given all available information [3]. Therefore,* (4.29) *can be written as*

$$\pi_k^*(\mathbb{I}_k): \ \tilde{z}_k^* = T_k^* \mathbb{E}[\tilde{e}_{k|k-1} | \mathbb{I}_k] + b_k, \ b_k \sim \mathcal{N}(0_m, \Theta).$$

This is the reason that $\pi_k^*(\mathbb{I}_k)$ *is called an information-based strategy. It is interesting to notice that the attacker needs to run another Kalman filter in order to compromise the existing one; the design of optimal attack policy is recast as an optimal state estimation problem. The optimality of the proposed attack is guaranteed by the optimality of MMSE estimation.*

Remark 4.6. *Comparing* (4.29) *with the linear attack model in the pioneering work [22, 23], it can be found that in order to maximize* $\mathrm{Tr}(\tilde{P}_{k|k})$*, the attacker* *should use a linear function of ξ_k, namely, the "best guess" of the current prediction error, to design the compromised innovation.*

Note that P_k^ξ in (4.25) depends explicitly on $\tilde{P}_{k|k-1}$, which by (4.26) is determined by $\tilde{P}_{k-1|k-1}$. To make the calculation a closed loop, $\tilde{P}_{k|k}$ should be evaluated at each instant. Substituting $\tilde{z}_k^*$ into (4.27), we have

$$\tilde{P}_{k|k} = A\tilde{P}_{k-1|k-1}A^{\mathrm{T}} + Q + K\Sigma K^{\mathrm{T}} - KT_k^* P_k^\xi - P_k^\xi (T_k^*)^{\mathrm{T}} K^{\mathrm{T}}. \tag{4.41}$$

The penultimate term of (4.41) satisfies

$$\begin{aligned} KT_k^* P_k^\xi &= -U\hat{S}\hat{\Sigma}^{\frac{1}{2}} V_k U_k^{\mathrm{T}} \hat{P}_k^{-\frac{1}{2}} U^{\mathrm{T}} P_k^\xi \\ &\overset{(a)}{=} -U\hat{P}_k^{-\frac{1}{2}} U_k S_k V_k^T V_k U_k^{\mathrm{T}} \hat{P}_k^{-\frac{1}{2}} U^{\mathrm{T}} P_k^\xi \\ &\overset{(b)}{=} -U\hat{P}_k^{-\frac{1}{2}} (\hat{P}_k^{\frac{1}{2}} \hat{S}\hat{\Sigma}\hat{S}\hat{P}_k^{\frac{1}{2}})^{\frac{1}{2}} \hat{P}_k^{-\frac{1}{2}} U^{\mathrm{T}} P_k^\xi \end{aligned} \tag{4.42}$$

where in (a) and (b) we use respectively the following equalities:

$$\hat{S}\hat{\Sigma}^{\frac{1}{2}} = \hat{P}_k^{-\frac{1}{2}} U_k S_k V_k^{\mathrm{T}},\ (\hat{P}_k^{\frac{1}{2}} \hat{S}\hat{\Sigma}\hat{S}\hat{P}_k^{\frac{1}{2}})^{\frac{1}{2}} = U_k S_k U_k^{\mathrm{T}}.$$

With (4.41)–(4.42), the attack performance can be evaluated without solving T_k^* explicitly. If A is stable, $\tilde{P}_{k|k}$ will converge to a constant. For a scalar system ($m = n = 1$), the performance evolution becomes

$$\tilde{P}_{i|i} = a^2 \tilde{P}_{i-1|i-1} + q + k^2\sigma + 2\sqrt{k^2\sigma(a^2\tilde{P}_{i-1|i-1} + q - P_i^e)}. \tag{4.43}$$

To avoid ambiguity, we use i to denote the time index and symbols with lower cases to represent the corresponding constant parameters. The recursive attack policy is summarized in Algorithm 4.1. In real attack scenarios, the computational complexity to obtain the optimal attack sequence should also be one of the attacker's main concerns. Yet in this work, the computational burden to design $\tilde{y}_k$ is not a serious issue. Because all involved matrices in Algorithm 4.1 are independent of measurement data and can be calculated offline. Theorem 4.1 involves only SVD and matrix multiplications; both of them are computationally efficient.

4.2.3 Separation Principle

The attack policy in Theorem 4.1 can be determined with two steps, leading to the so-called "separation principle". First, the MMSE estimate for $\tilde{e}_{k|k-1}$ is obtained based on available information; then, the following attack strategy is adopted:

$$\tilde{z}_k = T_k \xi_k + b_k,\ b_k \sim \mathcal{N}(0, \Theta_k). \tag{4.44}$$

Substituting this general attack model to $\mathbf{P}_{4.2}$, the explicit optimal solutions for T_k and Θ_k can be derived.

Remark 4.7. *It is interesting to notice that the investigated problem in this work bears some similarities to LQG control [91]. According to* (4.14), (4.17) *and* $\boldsymbol{P}_{4.2}$*, we can treat* $\tilde{z}_k$ *as the input of the linear time-invariant system. The control objective is to minimize the performance index, which measures the weighted correlation between system states and inputs. The differences from LQG control are that there is an additional constraint restricting the covariance of the input, and the one-step performance function without quadratic control and state costs is considered. Despite the differences, the well-known separation principle in stochastic control still holds. The optimal input is a combination of MMSE estimation and linear transformation. The controller gain in LQG is obtained from difference Riccati equations, whereas in this work the gain is derived by solving SDPs. Moreover, since* T_k^* *is determined by* P_k^ξ*, which depends on the estimation error covariance for* $\tilde{e}_{k|k-1}$*, it is clear that* T_k^* *depends implicitly on* $K_i^\xi, i \in [\![\bar{k}, k-1]\!]$*. It reveals another difference with LQG control, where the recursions determining controller and estimator gains are completely decoupled.*

4.2.4 Dynamic Linear Attack Model

In existing studies [22–24, 36, 38, 39, 42, 68, 84, 97], the attack is assumed to have a static linear model. For comparison, Theorem 4.1 shows that the

Algorithm 4.1 Optimal Attacks with Strict Stealthiness

```
Require: $x_{\bar{k}|\bar{k}-1}$ and online measurement $y_k, \hat{y}_k$.
Ensure: Optimal compromised measurement $\tilde{y}_k^*$.
Initialization: Set $\bar{\xi}_{\bar{k}} = 0_m, \bar{P}_{\bar{k}}^e = \bar{P}, \tilde{P}_{\bar{k}|\bar{k}-1} = \bar{P}$.
Do SVD: $K = U\hat{S}V^{\mathrm{T}}$. Calculate $G$ and $\Theta$.
Calculate $r_{\bar{k}} = \bar{y}_{\bar{k}} - \bar{C}x_{\bar{k}|\bar{k}-1}$.
for $k = \bar{k} : 1 : \infty$ do
    Calculate $K_k^{\xi}, \xi_k$ and $P_k^e$ with (4.20), (4.19) and (4.22).
    Calculate $P_k^{\xi}$ from (4.25).
    Design $T_k^*$ and $b_k$ according to Theorem 4.1.
    Calculate $\tilde{z}_k^*$ and $\tilde{y}_k^*$ with (4.29) and (4.28).
    Evaluate $\tilde{P}_{k|k}$ with (4.41).
```

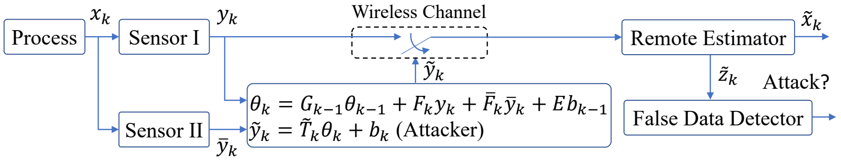

Figure 4.2: Deception attacks generated by a LTV system.

optimal attack has a dynamic linear form, owing to the fact that Kalman filter is a dynamic linear system. To show this, from (4.13) and (4.29), we obtain

$$\tilde{x}_{k|k-1} = A\tilde{x}_{k-1|k-2} + AKT_{k-1}^*\xi_{k-1} + AKb_{k-1}. \tag{4.45}$$

According to (4.18)–(4.19), we have

$$\xi_k = (I - K_k^{\xi}\bar{C})(A\xi_{k-1} - AK\tilde{z}_{k-1}^*) + K_k^{\xi}r_k. \tag{4.46}$$

Substitute (4.45) into (4.16), then it can be derived from (4.29) and (4.46) that

$$\begin{aligned}\xi_k = {} & (A - K_k^{\xi}\bar{C}A - AKT_{k-1}^*)\xi_{k-1} \\ & - K_k^{\xi}\bar{C}A\tilde{x}_{k-1|k-2} + K_k^{\xi}r_k - AKb_{k-1}.\end{aligned} \tag{4.47}$$

Consider the partitioned matrix $K^{\xi}_{k} = \left[K^{1}_{k}, K^{2}_{k}\right]$, where $K^{1}_{k} \in \mathbb{R}^{n\times m}, K^{2}_{k} \in \mathbb{R}^{n\times \bar{m}}$, and define the state vector $\theta_k \in \mathbb{R}^{2n}$ and matrices

$$\theta_k = \begin{bmatrix} \xi_k \\ \tilde{x}_{k|k-1} \end{bmatrix}, \ G_{k-1} = \begin{bmatrix} A - K^{\xi}_{k}\bar{C}A - AKT^{*}_{k-1} & -K^{\xi}_{k}\bar{C}A \\ AKT^{*}_{k-1} & A \end{bmatrix}$$

$$F_k = \begin{bmatrix} K^{1}_{k} \\ 0 \end{bmatrix}, \hat{F}_k = \begin{bmatrix} K^{2}_{k} \\ 0 \end{bmatrix}, E = \begin{bmatrix} -AK \\ AK \end{bmatrix}, \tilde{T}_k = \begin{bmatrix} T^{*}_{k} & C \end{bmatrix}.$$

From (4.28), (4.45) and (4.47), one can verify that $\tilde{y}_k$ is generated by the LTV system

$$\theta_k = G_{k-1}\theta_{k-1} + F_k y_k + \hat{F}_k \hat{y}_k + E b_{k-1} \tag{4.48}$$

$$\tilde{y}^{*}_{k} = \tilde{T}_k \theta_k + b_k \tag{4.49}$$

with initial condition $\theta_{\bar{k}} = \left[(K^{\xi}_{\bar{k}} r_{\bar{k}})^{\mathrm{T}}, x^{\mathrm{T}}_{\bar{k}|\bar{k}-1}\right]^{\mathrm{T}}$. b_k is an i.i.d. Gaussian noise. Note that

$$\begin{bmatrix} I & I \\ 0 & I \end{bmatrix} G_{k-1} \begin{bmatrix} I & I \\ 0 & I \end{bmatrix}^{-1} = \begin{bmatrix} A - K^{\xi}_{k}\bar{C}A & 0 \\ AKT^{*}_{k-1} & A - AKT^{*}_{k-1} \end{bmatrix}.$$

The eigenvalues of G_{k-1} consist of eigenvalues of $A - K^{\xi}_{k}\bar{C}A$ and $A - AKT^{*}_{k-1}$. This shows that the filter gain K^{ξ}_{k} and controller gain T^{*}_{k} can be designed separately, which coincides with Remark 4.7. The diagram of stealthy deception attacks against remote state estimation is illustrated in Fig. 4.2. In practical cases, to reduce the online computational burden, the attacker can calculate the coefficient matrices of the LTV system offline with Algorithm 4.1, then generate the optimal compromised measurement with (4.48)–(4.49). The online computational cost is negligible.

A similar problem to $\mathbf{P}_{4.2}$ was studied in [68], where it was assumed that besides the current (nominal) innovation, a fixed interval of historical innovations could be used. The attack has the following static linear form:

$$\tilde{z}_k = \sum_{i=0}^{\tau_k} T^{[i]}_{k} z_{k-i} + b_k, \ b_k \sim \mathcal{N}(0, \Phi_k) \tag{4.50}$$

where the interval length $\tau_k = \min\{k - \bar{k}, \tau\}$ is fixed for large k. With this assumption, $\mathbf{P}_{4.2}$ reduces to a convex optimization problem‡. From (4.50), we

‡The optimal solutions for $T^{[i]}_{k}, i \in [\![0, \tau_k]\!]$ and Φ_k are given in [68, Th. 1].

see that increasing the length of the historical interval *always* leads to better attack performance, because we have more degrees of freedom to design $\tilde{z}_k$. Ideally, letting $\tau_k = k - \bar{k}$, namely, using all historical innovations, achieves the maximum attack performance. However, in this case there will be a growing number of decision variables ($T_k^{[i]}$) as k increases, making the calculation intractable. Note that in [68] the attacker can only intercept the original measurement. In the remainder of this section we show that the attack in (4.29) with $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k\}$ and the optimal one in [68, Th. 1] utilizing all historical innovations have the same attack performance. If $\text{rank}(C) = m$, these two policies lead to the same $\tilde{z}_k$, and consequently, the same $\tilde{y}_k$.

Proposition 4.1. *The attack in* (4.29) *with* $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k\}$ *is equivalent to the optimal attack in* (4.50) *with* $\tau_k = k - \bar{k}$.

Proof. See Appendix A.2. ■

Remark 4.8. *Proposition 4.1 shows the connection between the proposed strategy and that in [68]. Two different methods lead to the same optimal policy when* $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k\}$ *and* $\tau_k = k - \bar{k}$*, but the one using static linear combinations becomes intractable as k tends to infinity. The proposed method in this paper is a recursive policy that has a simple form and is computationally efficient.*

4.3 Attacks with Relaxed Stealthiness

In this section, we study attacks with relaxed stealthiness. Let $\delta > 0$; the compromised innovation satisfies (4.9). Recall that $K = U\hat{S}V^{\mathrm{T}}$; we define the following variables:

$$\hat{K} = \hat{S}V^{\mathrm{T}},\ \hat{e}_k = U^{\mathrm{T}}\tilde{e}_{k|k-1},\ \hat{\xi}_k = U^{\mathrm{T}}\xi_k,\ \hat{P}_k = U^{\mathrm{T}}P_k^{\xi}U.$$

According to (4.27), $\mathbf{P}_{4.1}$ can be reformulated as

$$\mathbf{P}_{4.4}: \min_{\tilde{z}_k=\pi_k(\mathbb{I}_k)} \quad \mathrm{Tr}\{-\hat{K}\mathbb{E}[\tilde{z}_k\tilde{z}_k^{\mathrm{T}}]\hat{K}^{\mathrm{T}} + 2\mathbb{E}[\hat{e}_k\tilde{z}_k^{\mathrm{T}}]\hat{K}^{\mathrm{T}}\} \ \text{ s.t. } \ (4.9).$$

To solve $\mathbf{P}_{4.4}$, we first study the following MMSE estimate based linear attack and then prove its optimality.

$$\tilde{z}_k = T_k\hat{\xi}_k + b_k,\ b_k \sim \mathcal{N}(\beta_k, \Theta_k). \tag{4.51}$$

4.3.1 Linear Attack Based on MMSE Estimate

For brevity, denote $\hat{\mu}_k = \mathbb{E}[\hat{\xi}_k], \mu_k = \mathbb{E}[\tilde{z}_k], \Sigma_k = \mathrm{Cov}[\tilde{z}_k]$ and $\hat{\Sigma}_k = \mathbb{E}[\tilde{z}_k\tilde{z}_k^{\mathrm{T}}]$. Substitute (4.51) into $\mathbf{P}_{4.4}$, then the objective function becomes

$$f_k(\hat{\Sigma}_k, T_k, \beta_k) = \mathrm{Tr}[-\hat{K}\hat{\Sigma}_k\hat{K}^{\mathrm{T}} + 2(\hat{P}_kT_k^{\mathrm{T}} + \hat{\mu}_k\beta_k^{\mathrm{T}})\hat{K}^{\mathrm{T}}].$$

By linearity, $\tilde{z}_k$ is Gaussian distributed. The KL divergence between $\tilde{z}_k$ and z_k is given as

$$\mathcal{D}_{\mathrm{KL}}(\tilde{z}_k\|z_k) = \frac{1}{2}\left[\mathrm{Tr}(\Sigma^{-1}\Sigma_k) + \mu_k^{\mathrm{T}}\Sigma^{-1}\mu_k - m + \ln\frac{|\Sigma|}{|\Sigma_k|}\right].$$

To obtain the optimal linear attack, the attacker needs to solve the following optimization problem at each instant to obtain T_k^*, β_k^* and Θ_k^*:

$$\begin{aligned}
\mathbf{P}_{4.5}: \min_{\mu_k,\Sigma_k,\hat{\Sigma}_k,T_k,\beta_k,\Theta_k} \quad & f_k(\hat{\Sigma}_k, T_k, \beta_k) \\
\text{s.t.} \quad & T_k[\hat{P}_k - \hat{\mu}_k\hat{\mu}_k^{\mathrm{T}}]T_k^{\mathrm{T}} + \Theta_k = \Sigma_k && (4.52)\\
& \Sigma_k + \mu_k\mu_k^{\mathrm{T}} = \hat{\Sigma}_k && (4.53)\\
& T_k\hat{\mu}_k + \beta_k = \mu_k && (4.54)\\
& g(\Sigma_k, \hat{\Sigma}_k) \le 0 && (4.55)\\
& \Theta_k \succeq 0 && (4.56)
\end{aligned}$$

where (4.55) is the stealthiness constraint with $g(\cdot)$ defined by

$$g(\Sigma_k, \hat{\Sigma}_k) = \mathrm{Tr}(\Sigma^{-1}\hat{\Sigma}_k) + \ln\frac{|\Sigma|}{|\Sigma_k|} - m - 2\sigma.$$

The independent variables in $\mathbf{P}_{4.5}$ are $\hat{\Sigma}_k, T_k$ and β_k. At the kth sampling instant, $\hat{\mu}_k = U^{\mathrm{T}}\mathbb{E}[\xi_k]$. Since the MMSE estimation is unbiased, i.e., $\mathbb{E}[\xi_k] = \mathbb{E}[\tilde{e}_{k|k-1}]$, according to (4.14), we have

$$\mathbb{E}[\xi_k] = A\mathbb{E}[\xi_{k-1}] - AK\mu_{k-1} \tag{4.57}$$

with initial condition $\mathbb{E}[\xi_{\bar{k}}] = 0$. This equality is used recursively to determine $\hat{\mu}_k$. $\hat{P}_k$ denotes the second moment of $\hat{\xi}_k$ and is obtain from (4.25). Hence, $\hat{\mu}_k$ and $\hat{P}_k$ are constant parameters at the kth instant.

In (4.51), b_k serves as a compensation term to ensure stealthy. Contrary to the case of strict stealthiness, it is assumed that b_k may have non-zero and time-varying mean. With this model, $\tilde{z}_k$ can have an arbitrary Gaussian distribution. In the following proposition, we show that the optimal compromised innovation should have a compensation term with zero mean, which can reduce the complexity of $\mathbf{P}_{4.5}$ and also provide additional insights on the attacker's optimal behavior.

Proposition 4.2. *The optimal attack based on MMSE estimate in* (4.51) *with relaxed stealthiness satisfies*

$$\beta_k = 0, \ \forall k \geq \bar{k}. \tag{4.58}$$

If $\operatorname{rank}(C) = m$*, then* $b_k = 0$*.*

Proof. See Appendix A.3. ∎

Now $\mathbf{P}_{4.5}$ reduces to

$$\begin{aligned} \mathbf{P}_{4.6}: \quad \min_{T_k, \Sigma_k} \quad & \operatorname{Tr}(-\hat{K}\Sigma_k\hat{K}^{\mathrm{T}} + 2\hat{K}T_k\hat{P}_k) \\ \text{s.t.} \quad & \operatorname{Tr}(\Sigma^{-1}\Sigma_k) + \ln\frac{|\Sigma|}{|\Sigma_k|} - m - 2\delta \leq 0 \\ & T_k\hat{P}_kT_k^{\mathrm{T}} - \Sigma_k \preceq 0. \end{aligned}$$

The first constraint is convex in Σ_k and the second one can be reformed as a linear matrix inequality by applying Schur complement. $\mathbf{P}_{4.6}$ is convex and can be solved efficiently.

In this section, we have derived the optimal linear attack based on MMSE estimate with relaxed stealthiness. The linear model preserves Gaussianity; as a result, the compromised innovation is Gaussian distributed. Both the objective function and stealthiness constraints of $\mathbf{P}_{4.4}$ have analytical expressions, which enables derivation of attack coefficients by solving convex optimization problems. However, the constraint in (4.9) does not require that $\tilde{z}_k$ be Gaussian. In the following section, we will show that the MMSE estimate based linear attack is optimal among all feasible $\tilde{z}_k$, which can have an arbitrary probability distribution.

4.3.2 Optimal Attack Policy

With ξ_k and P_k^{ξ} obtained from (4.19) and (4.25), we have the following result.

Theorem 4.2. *If $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k, \hat{y}_k\}$, or $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k\}$, the optimal attack policy with relaxed stealthiness is given by*

$$\pi_k^*(\mathbb{I}_k): \ \tilde{y}_k^* = C\tilde{x}_{k|k-1} + T_k^* U^{\mathrm{T}} \xi_k + b_k, \ b_k \sim \mathcal{N}(0, \Theta_k) \tag{4.59}$$

where T_k^* is obtained by solving $\boldsymbol{P}_{4.6}$, $\Theta_k = \Sigma_k^* - T_k^* \hat{P}_k (T_k^*)^{\mathrm{T}}$.

Proof. According to (4.59), the optimal compromised innovation is

$$\tilde{z}_k^* = T_k^* \hat{\xi}_k + b_k, \ b_k \sim \mathcal{N}(0, \Theta_k).$$

At the $\bar{k}$th sampling instant, assume that $\tilde{z}_{\bar{k}}^a = \pi_{\bar{k}}(\mathbb{I}_{\bar{k}})$ is an arbitrary attack policy (not necessarily Gaussian). Denote

$$\mathbb{E}[\tilde{z}_{\bar{k}}^a] = \tilde{\mu}_{\bar{k}}, \ \mathbb{E}[(\tilde{z}_{\bar{k}}^a - \tilde{\mu}_{\bar{k}})(\tilde{z}_{\bar{k}}^a - \tilde{\mu}_{\bar{k}})^{\mathrm{T}}] = \tilde{\Sigma}_{\bar{k}}.$$

Consider the following optimization problem:

$$\min_{\tilde{z}_{\bar{k}} = \pi_{\bar{k}}(\mathbb{I}_{\bar{k}})} \ \mathrm{Tr}\{\hat{K}\mathbb{E}[\tilde{z}_{\bar{k}} \hat{e}_{\bar{k}}^{\mathrm{T}}]\} \quad \text{s.t.} \ \ \tilde{z}_{\bar{k}} \sim \mathcal{N}(\tilde{\mu}_{\bar{k}}, \tilde{\Sigma}_{\bar{k}}). \tag{4.60}$$

The optimal policy obtained from (4.60) is denoted as $\tilde{z}_{\bar{k}}^b$. Note that (4.60) is a similar problem to (4.30). Following the proof of Theorem 4.1, we see that $\tilde{z}_{\bar{k}}^b$ has the form

$$\tilde{z}_{\bar{k}}^b = \tilde{T}_{\bar{k}}^* \hat{\xi}_{\bar{k}} + \tilde{b}_{\bar{k}}, \ \tilde{b}_{\bar{k}} \sim \mathcal{N}(\tilde{\mu}_{\bar{k}}, \tilde{\Theta}_{\bar{k}})$$

where $\tilde{T}_{\bar{k}}^* \hat{P}_{\bar{k}} (\tilde{T}_{\bar{k}}^*)^{\mathrm{T}} + \tilde{\Theta}_{\bar{k}} = \tilde{\Sigma}_{\bar{k}}$. Since $\tilde{z}_{\bar{k}}^a$ and $\tilde{z}_{\bar{k}}^b$ have the same second moment, according to the proof of Theorem 4.1, it can be verified that the objective value of $\tilde{z}^b_{\bar{k}}$ for $\mathbf{P}_{4.4}$ is no greater than that of $\tilde{z}^a_{\bar{k}}$. Let $p_{\tilde{z}^a_{\bar{k}}}(t)$ and $p_{\tilde{z}^b_{\bar{k}}}(t)$ denote the probability density functions of $\tilde{z}^a_{\bar{k}}$ and $\tilde{z}^b_{\bar{k}}$, respectively. Note that

$$\int_{\mathbb{R}^m} p_{\tilde{z}^a_{\bar{k}}}(t) t^{\mathrm{T}} \Sigma^{-1} t \mathrm{d}t = \mathrm{Tr}(\Sigma^{-1}\mathbb{E}[\tilde{z}^a_{\bar{k}}(\tilde{z}^a_{\bar{k}})^{\mathrm{T}}])$$

$$\int_{\mathbb{R}^m} p_{\tilde{z}^b_{\bar{k}}}(t) t^{\mathrm{T}} \Sigma^{-1} t \mathrm{d}t = \mathrm{Tr}(\Sigma^{-1}\mathbb{E}[\tilde{z}^b_{\bar{k}}(\tilde{z}^b_{\bar{k}})^{\mathrm{T}}]).$$

It follows that

$$\begin{aligned}
&\mathcal{D}_{\mathrm{KL}}(\tilde{z}^b_{\bar{k}} \| z_{\bar{k}}) - \mathcal{D}_{\mathrm{KL}}(\tilde{z}^a_{\bar{k}} \| z_{\bar{k}}) \\
=& \int_{\mathbb{R}^m} p_{\tilde{z}^b_{\bar{k}}}(t) \ln \frac{p_{\tilde{z}^b_{\bar{k}}}(t)}{p_{z_{\bar{k}}}(t)} \mathrm{d}t - \int_{\mathbb{R}^m} p_{\tilde{z}^a_{\bar{k}}}(t) \ln \frac{p_{\tilde{z}^a_{\bar{k}}}(t)}{p_{z_{\bar{k}}}(t)} \mathrm{d}t \\
=& \{-\mathcal{E}_{\tilde{z}^b_{\bar{k}}} + \frac{1}{2}\ln[(2\pi)^m|\Sigma|] + \frac{1}{2}\int_{\mathbb{R}^m} p_{\tilde{z}^b_{\bar{k}}}(t) t^{\mathrm{T}}\Sigma^{-1} t \mathrm{d}t\} - \\
& \{-\mathcal{E}_{\tilde{z}^a_{\bar{k}}} + \frac{1}{2}\ln[(2\pi)^m|\Sigma|] + \frac{1}{2}\int_{\mathbb{R}^m} p_{\tilde{z}^a_{\bar{k}}}(t) t^{\mathrm{T}}\Sigma^{-1} t \mathrm{d}t\} \\
=& \mathcal{E}_{\tilde{z}^a_{\bar{k}}} - \mathcal{E}_{\tilde{z}^b_{\bar{k}}} \leq 0.
\end{aligned}$$

The last inequality is due to the fact that Gaussian distribution has the maximum entropy among all probability distributions with the same variance [12, 23]. It shows that for any feasible policy $\tilde{z}^a_{\bar{k}}$, we can always find another MMSE estimate based policy $\tilde{z}^b_{\bar{k}}$, such that $\tilde{z}^b_{\bar{k}}$ can cause no less performance loss and also satisfy the stealthiness constraint. Since $\pi_k(\mathbb{I}_k)$ is a recursive strategy, following the same arguments it can be verified that the conclusion holds $\forall k \geq \bar{k}$. The above analysis proves the optimality of $\tilde{z}^*_k$. ■

The algorithm for designing deception attacks with relaxed stealthiness is similar to Algorithm 4.1, where T^*_k and Θ_k are obtained by solving $\mathbf{P}_{4.6}$, and the attack performance is evaluated by

$$\tilde{P}_{k|k} = A\tilde{P}_{k-1|k-1}A^{\mathrm{T}} + Q + K\Sigma^*_k K^{\mathrm{T}} - KT^*_k P^{\xi}_k - P^{\xi}_k (T^*_k)^{\mathrm{T}} K^{\mathrm{T}}. \tag{4.61}$$

Remark 4.9. *Theorem 4.2 shows that the separation principle still holds for designing attacks with relaxed stealthiness, i.e., the attacker first utilizes all available information to obtain an MMSE estimate for $\tilde{e}_{k|k-1}$, then the optimal transformation matrix in* (4.59) *is derived by solving convex optimization* *problems. Consequently, the optimal attack policy can also be generated by the LTV system in* (4.48)–(4.49).

Remark 4.10. *With the optimal attack based on MMSE estimate, Proposition 4.2 leads to* $\mathbb{E}[\tilde{e}_{k|k-1}] = 0$ *and* $\mathbb{E}[\tilde{z}_k] = 0,\ \forall k \geq \bar{k}$*. This implies that the remote estimator still provides unbiased state estimation with compromised measurements. The performance degradation is due to the increase of estimation error covariance. The conclusion is different from [38], where the interval attack performance is considered and* $b_k \neq 0$ *for large* δ*, leading to biased state estimation.*

4.4 Performance Analysis and Comparison

In this section, we study further the optimal attacks with different information sets. The general scenario in (4.7) can be tackled by the separation principle. Some interesting results are obtained when we investigate the special cases. We consider only the attacks with strict stealthiness. The extension to the case of relaxed stealthiness is straightforward.

4.4.1 Null Information: $\mathbb{I}_k = \emptyset$

In practical cases, owing to some defensive countermeasures deployed by system operators, attackers may not be able to eavesdrop any measurement data. Then $\mathbb{I}_k = \emptyset, \forall k \geq \bar{k}$; attackers can still launch stealthy deception attacks if they can modify the transmitted packets. Note that it is equivalent to assuming $\lambda_{m+\bar{m}}(\tilde{R}) \to \infty$. From (4.18)–(4.22) we have $K_k^{\xi} = 0, \xi_k = \bar{\xi}_k$ and $P_k^e = \bar{P}_k^e$; then

$$\xi_k = A\xi_{k-1} - AK\tilde{z}_{k-1} \tag{4.62}$$

$$\bar{P}_k^e = A\bar{P}_{k-1}^e A^{\mathrm{T}} + Q. \tag{4.63}$$

In this case, Theorem 4.1 cannot be applied directly because $\hat{P}_k$ can be singular. For example, when attack starts, we have $\hat{P}_{\bar{k}} = U^{\mathrm{T}}(\tilde{P}_{\bar{k}|\bar{k}-1} - P_{\bar{k}}^e)U = 0$. The optimal attack policy is summarized in the following theorem.

Theorem 4.3. *If $\mathbb{I}_k = \emptyset$, the optimal attack policy with strict stealthiness is given by*

$$\pi_k^*(\mathbb{I}_k): \ \tilde{y}_k^* = C\tilde{x}_{k|k-1} + T_k^*\xi_k + b_k, \ b_k \sim \mathcal{N}(0, \Theta_k^*) \tag{4.64}$$

with the coefficients

$$T_k^* = -\Sigma^{\frac{1}{2}}\Phi_k\Pi_k^{-\frac{1}{2}}\Phi_k^{\mathrm{T}}\Sigma^{\frac{1}{2}}K^{\mathrm{T}}\Psi_k\Psi_k^{\mathrm{T}} + \Sigma^{\frac{1}{2}}\bar{\Phi}_k\mathcal{X}_k\Lambda_k^{-\frac{1}{2}}\Psi_k^{\mathrm{T}} + \mathcal{Y}_k\bar{\Psi}_k^{\mathrm{T}}$$

$$\Theta_k^* = \Sigma^{\frac{1}{2}}\bar{\Phi}_k(I - \mathcal{X}_k\mathcal{X}_k^{\mathrm{T}})\bar{\Phi}_k^{\mathrm{T}}\Sigma^{\frac{1}{2}}$$

where $\Phi_k, \bar{\Phi}_k, \Pi_k$ and $\Psi_k, \bar{\Psi}_k, \Lambda_k$ satisfy the spectral decompositions:

$$\Sigma^{\frac{1}{2}}K^{\mathrm{T}}P_k^{\xi}K\Sigma^{\frac{1}{2}} = \begin{bmatrix}\Phi_k & \bar{\Phi}_k\end{bmatrix}\begin{bmatrix}\Pi_k & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\Phi_k^{\mathrm{T}}\\ \bar{\Phi}_k^{\mathrm{T}}\end{bmatrix}, \ \Pi_k \in \mathbb{S}_{++}^{\bar{r}}$$

$$P_k^{\xi} = \begin{bmatrix}\Psi_k & \bar{\Psi}_k\end{bmatrix}\begin{bmatrix}\Lambda_k & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\Psi_k^{\mathrm{T}}\\ \bar{\Psi}_k^{\mathrm{T}}\end{bmatrix}, \ \Lambda_k \in \mathbb{S}_{++}^{s}$$

and $\mathcal{X}_k \in \mathbb{R}^{(m-\bar{r})\times s}, \mathcal{Y}_k \in \mathbb{R}^{m\times(n-s)}$ are free parameters satisfying

$$\mathcal{X}_k\mathcal{X}_k^{\mathrm{T}} \preceq I, \ \mathcal{X}_k\Lambda_k^{\frac{1}{2}}\Psi_k^{\mathrm{T}}K\Sigma^{\frac{1}{2}}\Phi_k = 0. \tag{4.65}$$

Proof. See Appendix A.4. ■

When $\mathbb{I}_k = \emptyset$, optimal attack policies are not unique even if $\mathrm{rank}(C) = m$. All policies with different feasible $\mathcal{X}_k$ and $\mathcal{Y}_k$ are stealthy and have the same performance, but the effects of these free parameters are different. $\mathcal{Y}_k$ exists because P_k^{ξ} can be singular. The last term of T_k^* does not influence the constraint in (A.31) and satisfies $\mathcal{Y}_k\bar{\Psi}_k^{\mathrm{T}}P_k^{\xi}K = 0$, indicating that different values of $\mathcal{Y}_k$ contribute the same change to $\tilde{P}_{k|k}$, not only the same increase to $\mathrm{Tr}(\tilde{P}_{k|k})$. On the contrary, $\mathcal{X}_k$ has an impact on the stealthiness constraint; hence, Θ_k^* serves as a compensation term to ensure stealthiness. Different selections of $\mathcal{X}_k$ result in different changes in $\tilde{P}_{k|k}$, which, by the recursion in (4.26)–(4.27), will influence the attack performance in the subsequent steps. In this paper we consider only the greedy attack policy, which maximizes $\mathrm{Tr}(\tilde{P}_{k|k})$ given $\tilde{P}_{k-1|k-1}$; hence, $\mathcal{X}_k$ and $\mathcal{Y}_k$ can be chosen freely. Let $\mathcal{X}_k = 0$ and $\mathcal{Y}_k = -\Sigma^{\frac{1}{2}}\Phi_k\Pi_k^{-\frac{1}{2}}\Phi_k^{\mathrm{T}}\Sigma^{\frac{1}{2}}K^{\mathrm{T}}\bar{\Psi}_k$; an optimal strategy with a simple form is

$$T_k^* = -\Sigma^{\frac{1}{2}}\Phi_k\Pi_k^{-\frac{1}{2}}\Phi_k^{\mathrm{T}}\Sigma^{\frac{1}{2}}K^{\mathrm{T}}, \ \Theta_k^* = \Sigma^{\frac{1}{2}}\bar{\Phi}_k\bar{\Phi}_k^{\mathrm{T}}\Sigma^{\frac{1}{2}}.$$

At the $\bar{k}$th sampling instant, $\xi_{\bar{k}} = 0, P^{\xi}_{\bar{k}} = 0$. It can be verified that $T^*_{\bar{k}}$ is an arbitrary matrix and $\Theta^*_{\bar{k}} = \Sigma$. The compromised innovation satisfies $\tilde{z}^*_{\bar{k}} \sim \mathcal{N}(0, \Sigma)$. The algorithm for designing $\tilde{z}_k$ is the same as Algorithm 4.1, where T^*_k and Θ^*_k are obtained from Theorem 4.3. The attack performance can be evaluated by (4.41).

Remark 4.11. *If $\mathbb{I}_k = \emptyset$, one feasible attack policy is designing $\tilde{z}_k$ as a white Gaussian noise, i.e., $\tilde{z}_k \sim \mathcal{N}(0, \Sigma), \mathbb{E}[\tilde{z}_i \tilde{z}_j^{\mathrm{T}}] = 0, \forall i \neq j$. In this case the a posteriori estimation error covariance evolves according to*

$$\tilde{P}_{k|k} = A\tilde{P}_{k-1|k-1}A^{\mathrm{T}} + Q + K\Sigma K^{\mathrm{T}}. \tag{4.66}$$

Compared with this intuitive strategy, one can verify that (4.64) *causes more severe performance loss. This is because the attacker makes full utilization of the information contained in $\{\tilde{z}_{\bar{k}}, ..., \tilde{z}_{k-1}\}$ to design $\tilde{z}_k$. Note that at the $\bar{k}$th sampling instant, these two attacks are identical (in the sense of leading to the same attack performance).*

4.4.2 Symmetric Information: $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k\}$

A majority of existing studies concentrates on the scenario that the attacker can only eavesdrop on the original measurement. In this case, $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k\}$ is the symmetric information that is available to both the attacker and remote estimator.

Since $\bar{C} = C, \bar{R} = R, \bar{P}^e_{\bar{k}} = \bar{P}$, (4.18)–(4.22) is a steady-state Kalman filter, leading to $K^{\xi}_k = K,\ P^e_k = (I - KC)\bar{P},\ \forall k \geq \bar{k}$. With fixed P^e_k, P^{ξ}_k is determined only by $\tilde{P}_{k|k-1}$; thus Algorithm 4.1 can be simplified. The following proposition reveals the connection between policy (4.29) and the conclusion in [22].

Proposition 4.3. *If $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k\}$, the attack in* (4.29) *satisfies*

$$K(\tilde{z}^*_{\bar{k}} + z_{\bar{k}}) = 0. \tag{4.67}$$

If $\operatorname{rank}(C) = m$*, then $\tilde{z}^*_{\bar{k}} = -z_{\bar{k}}$.*

Proof. See Appendix A.5. ■

From (4.67), one can verify that $\tilde{z}_{\bar{k}}^*$ from Theorem 4.1 and $\tilde{z}_{\bar{k}} = -z_{\bar{k}}$ have the same attack performance, indicating the equivalence between these two attacks. If C has full row rank, the unique attack policy is $\tilde{z}_{\bar{k}}^* = -z_{\bar{k}}$, which is consistent with the conclusion in [22]. When $k \geq \bar{k}+1$, more information can be used to obtain a more accurate estimate for $\tilde{e}_{k|k-1}$. This is the reason that the attack in (4.29) outperforms the one in [22] if only (4.10) serves as the stealthiness constraint§.

An interesting observation is that, in case of symmetric information and $\mathrm{rank}(C) = m$, if we just flip the sign of T_k^* in Algorithm 4.1, $\tilde{P}_{k|k}$ will eventually converge to $(I - KC)\bar{P}$; that is, the estimator becomes a nominal Kalman filter. Consider a similar optimization problem to $\mathbf{P}_{4.2}$ but with an opposite objective:

$$\min_{\tilde{z}_k = \pi_k(\mathbb{I}_k)} \quad -\mathrm{Tr}\{K\mathbb{E}[\tilde{z}_k \tilde{e}_{k|k-1}^{\mathrm{T}}]\} \quad \text{s.t.} \quad (4.10).$$

It is clear that the optimal solution to the above problem is $\tilde{z}_k = -T_k^* \xi_k$ and will result in a standard Kalman filer.

4.4.3 Full Information: $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k, \hat{y}_k\}$

The solution to the full information case is given in Theorem 4.1. The Kalman filter in (4.18)–(4.22) is time-varying and will eventually converge to a fixed-gain filter. The filter gain is

$$K^{\xi} = \lim_{k\to\infty} K_k^{\xi} = \bar{P}^e \bar{C}^T (\bar{C}\bar{P}^e\bar{C}^T + \bar{R})^{-1}$$

where $\bar{P}^e$ is the unique solution of the Riccati equation

$$\bar{P}^e = h[g_{[\bar{C},\bar{R}]}(\bar{P}^e)].$$

In the extreme case that the attacker has a very accurate measurement, i.e., $\hat{R} = 0$, (4.22) implies $P_k^e = 0$ and $P_k^{\xi} = \tilde{P}_{k|k-1}, \forall k \geq \bar{k}$; this leads to some simplification of Algorithm 4.1.

§In this paper, the χ^2 detector utilizes only the current $\tilde{z}_k$ to calculate the index function. It is not necessary to require that $\mathbb{E}[\tilde{z}_i \tilde{z}_j^{\mathrm{T}}] = 0, \forall i \neq j$. In this case, the MMSE estimate based attack outperforms other feasible attacks that satisfy the stealthiness constraint.

4.4.4 Performance Comparison

In Theorem 4.1, we see that the attack performance depends on the estimation quality for $\tilde{e}_{k|k-1}$. Additional information will *always* benefit the attacker's purpose. In this section, we use the scalar system to better explain this idea. Consider the following special cases for (4.43):

$$\tilde{P}_{i|i} = a^2\tilde{P}_{i-1|i-1} + q - k^2\sigma \tag{4.68}$$

$$\tilde{P}_{i|i} = a^2\tilde{P}_{i-1|i-1} + q \tag{4.69}$$

$$\tilde{P}_{i|i} = a^2\tilde{P}_{i-1|i-1} + q + k^2\sigma \tag{4.70}$$

$$\tilde{P}_{i|i} = a^2\tilde{P}_{i-1|i-1} + q + k^2\sigma + 2\sqrt{k^2\sigma(a^2\tilde{P}_{i-1|i-1} + q - \bar{P}^e_i)} \tag{4.71}$$

$$\tilde{P}_{i|i} = a^2\tilde{P}_{i-1|i-1} + q + k^2\sigma + 2\sqrt{k^2\sigma(a^2\tilde{P}_{i-1|i-1} + q - P^e)} \tag{4.72}$$

$$\tilde{P}_{i|i} = a^2\tilde{P}_{i-1|i-1} + q + k^2\sigma + 2\sqrt{k^2\sigma(a^2\tilde{P}_{i-1|i-1} + q - P^e_i)} \tag{4.73}$$

$$\tilde{P}_{i|i} = a^2\tilde{P}_{i-1|i-1} + q + k^2\sigma + 2\sqrt{k^2\sigma(a^2\tilde{P}_{i-1|i-1} + q)} \tag{4.74}$$

where (4.68) is the performance evolution when there is no deception attacks. It corresponds to a standard Kalman filter. We have $\tilde{P}_{i|i} = (1-kc)\bar{P}$, $\forall i \geq \bar{i}$. (4.69) is the case when the measurement data is unavailable for estimation update; this can be regarded as DoS attacks (the attack cannot remain stealthy). (4.70) corresponds to (4.66), where $\tilde{z}_i$ is a white Gaussian noise. (4.71) is the case when $\mathbb{I}_i = \emptyset$, where $\bar{P}^e_i$ is given recursively by (4.63). One can verify that $\bar{P}^e_i \geq (1-kc)\bar{P}, \forall i \geq \bar{i}$. (4.72) is the case when $\mathbb{I}_i = \mathbb{I}_{i-1} \cup \{y_i\}$. $P^e = (1-kc)\bar{P}$ is a constant. (4.73) denotes the general case when $\mathbb{I}_i = \mathbb{I}_{i-1} \cup \{y_i, \hat{y}_i\}$. In the proof of Lemma 4.3, we have shown that $P^e_i \leq (1-kc)\bar{P}, \forall i \geq \bar{i}$. (4.74) is the extreme case when $\hat{R} = 0$.

In (4.71)–(4.73), the attacker's information set becomes larger in sequence. We see that the first three terms of $\tilde{P}_{i|i}$ are constants, and the last term increases in turn. This implies that the additional information causes more severe estimation quality deterioration. Specifically, (4.74) corresponds to the maximum performance loss the attacker can expect. If $|a| < 1$, $\tilde{P}_{i|i}$ will converge to the fixed point of the nonlinear equation

$$\tilde{P} = (\sqrt{a^2\tilde{P} + q} + \sqrt{k^2\sigma})^2.$$

4.5 Examples

In this section, we use numerical examples to verify the optimality of the proposed attack policy. Consider a stable LTI system with the following parameters:

$$A = \begin{bmatrix} 0.482 & -0.134 & 0.037 \\ -0.061 & 0.572 & -0.061 \\ -0.109 & -0.029 & 0.446 \end{bmatrix}, Q = \operatorname{diag}\left\{ \begin{bmatrix} 0.612 \\ 0.435 \\ 0.754 \end{bmatrix} \right\}$$

$$C = \begin{bmatrix} 1.326 & 0.756 & 2.352 \\ -1.319 & 0.921 & 0.395 \end{bmatrix}, R = \operatorname{diag}\left\{ \begin{bmatrix} 1.054 \\ 2.026 \end{bmatrix} \right\}, \hat{R} = 2$$

$$\hat{C} = \begin{bmatrix} 0.505 & 1.214 & 1.984 \end{bmatrix},\ S^{\mathrm{T}} = \begin{bmatrix} 0.132 & -0.814 \end{bmatrix}.$$

Assume $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k\}$; the attack is launched from $\bar{k} = 301$; the attack performance with strict and relaxed stealthiness is illustrated in Fig. 4.3. The theoretical evolution of $\mathrm{Tr}(\tilde{P}_{k|k})$ is derived from (4.41) and (4.61); the empirical one is obtained by simulating (3.1)–(3.2) for 20,000 times with randomly generated noises and averaging the corresponding square-errors at each sampling instant. It can be observed that the proposed attack can degrade the estimation quality significantly. For the considered stable system, $\mathrm{Tr}(\tilde{P}_{k|k})$ converges to a constant. It is also interesting to notice that the innovation-based linear attack with relaxed stealthiness ($\delta = 0.1$) in [23] causes less estimation performance loss compared with the strictly stealthy attack in Theorem 4.1. Note that $m = 2$; by setting the threshold of χ^2 detector as 4, the theoretical FAR in the nominal condition is 13.5%. Fig. 4.4 shows the empirical FAR. It is clear that the strictly stealthy attack can completely bypass the false-data detector (the empirical alarm rates with and without attacks fluctuate in a narrow interval, and coincide with the theoretical one). The attack with relaxed stealthiness causes more severe performance loss, with the price of a higher alarm rate.

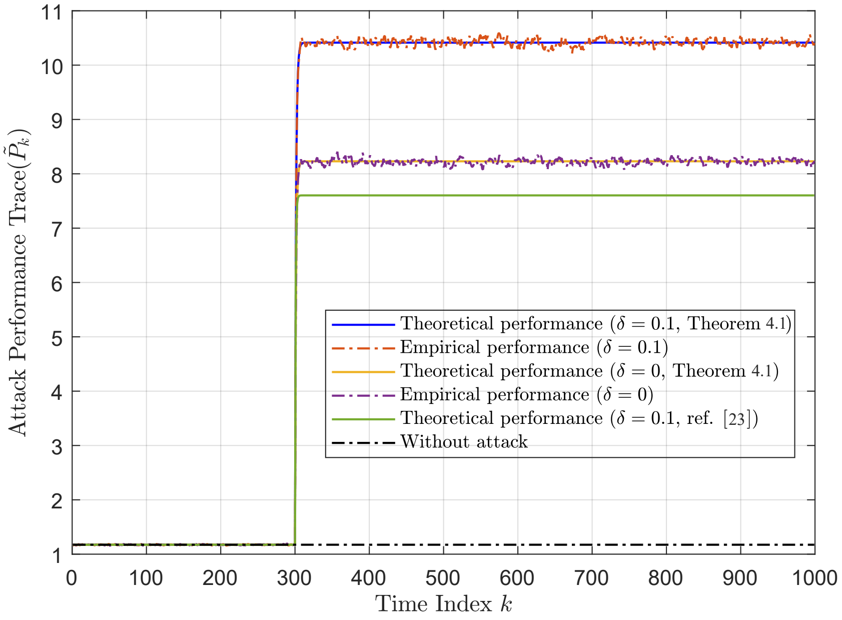

Figure 4.3: Attack performance with strict/relaxed stealthiness.

Assume $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k\}, \bar{k} = 16$. Fig. 4.5 compares the attack performance with symmetric information. It can be observed that the proposed information-based policy leads to more performance degradation compared with innovation-based ones. Specifically, the marked blue line illustrates the performance of the linear attack in (4.50) with interval length $\tau = 3$. When $16 \leq k \leq 19$, both policies of Theorem 4.1 and (4.50) are optimal, since all historical innovations are utilized. When $k \geq 20$, the innovations in $[\![16, k-4]\!]$ are out of the preset historical interval; hence (4.50) is no longer optimal and yields less estimation error. This is the reason that the two performance curves diverge apart when $k \geq 20$. Note that all optimal attacks with strict stealthiness have the same performance when $k = 16$. The simulation verifies Proposition 4.1.

Assume $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k, \hat{y}_k\}$. Fig. 4.6 illustrates the attack performance when the attacker has additional measurements. Compared with existing work, it shows that the information-based policy leads to greater performance loss. Let $S = 0$ and $\hat{R}$ vary from 0.001 to 1000. Fig. 4.7 shows the impact of

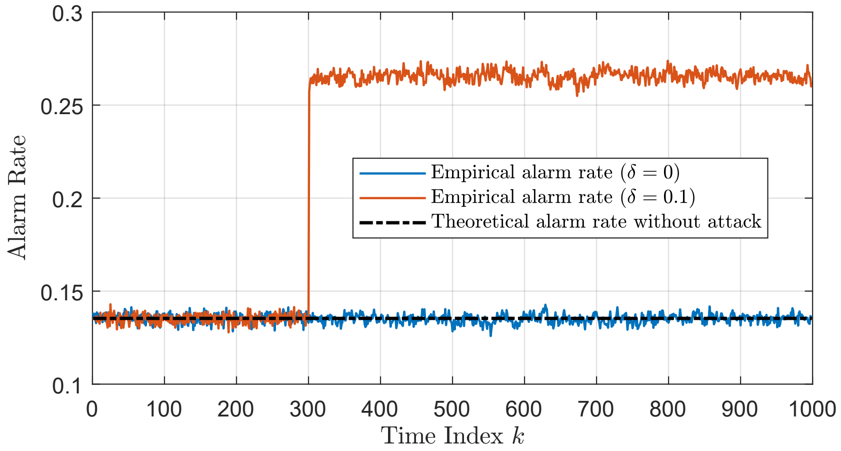

Figure 4.4: Empirical alarm rates with/without deception attacks.

side information on the steady-state attack performance. It is observed that a more accurate measurement of system states improves the attack performance.

Assume $\mathbb{I}_k = \emptyset$. Fig. 4.8 illustrates the attack performance when the attacker cannot eavesdrop any measurement data. Compared with the white noise attack in (4.66) and DoS attacks, the proposed strategy can cause more severe performance degradation. Note that the attacks from Theorem 4.3 and (4.66) have the same performance when they start ($k = 16$).

4.6 Conclusion

In this chapter, optimal deception attacks with both strict and relaxed stealthiness against remote state estimation are studied from an information perspective. General scenarios in which the attacker has different information sets from the remote estimator are investigated in a unified framework. Contrary to existing studies, we have designed an information-based stealthy attack policy that can cause the maximum performance loss. It is found that the attack performance depends highly on the estimation quality for the prediction error. The optimal compromised measurement is generated by an LTV system whose coefficient matrices can be determined without knowing the online measurements. The future work will be exploring the optimal information-based

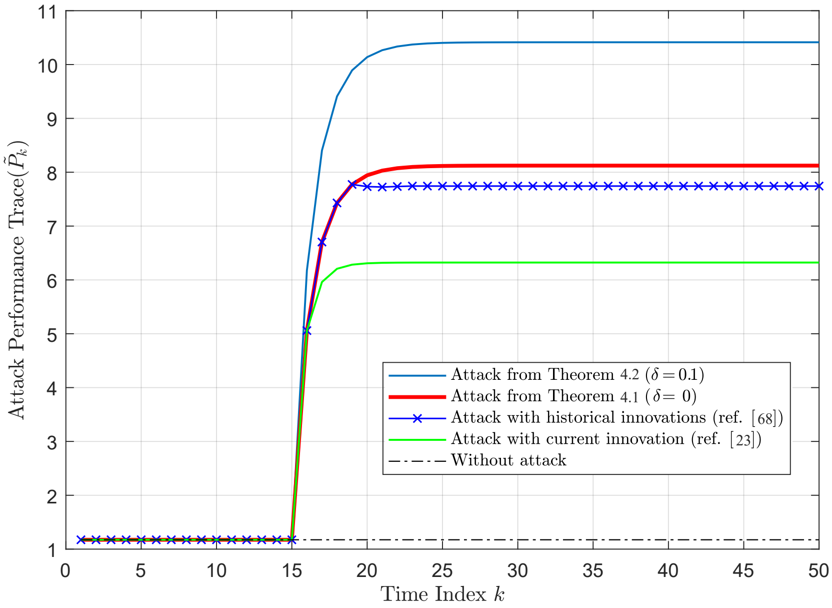

Figure 4.5: Attack performance with symmetric information.

stealthy attack considering the interval performance criterion, and studying the defensive countermeasures against deception attacks.

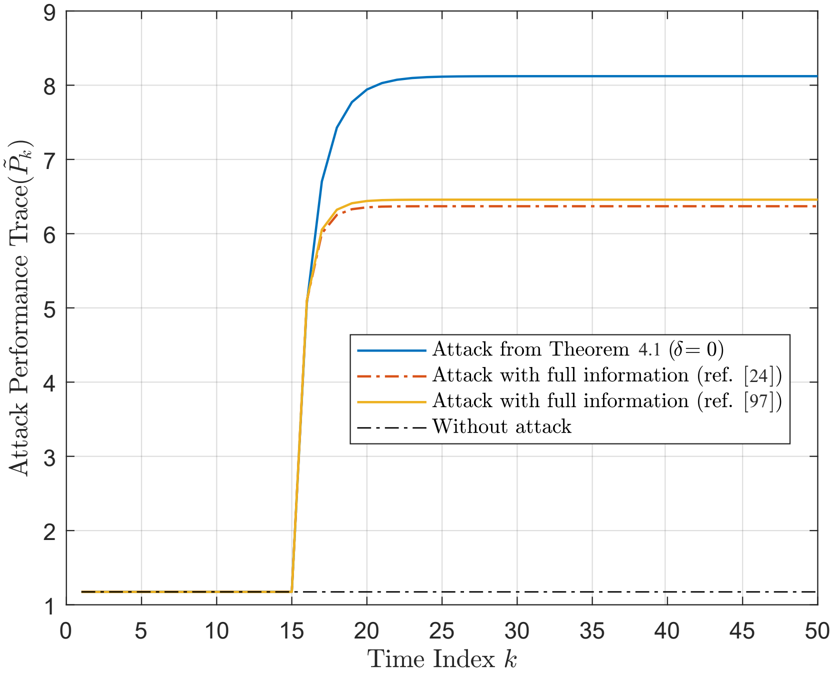

Figure 4.6: Attack performance with full information.

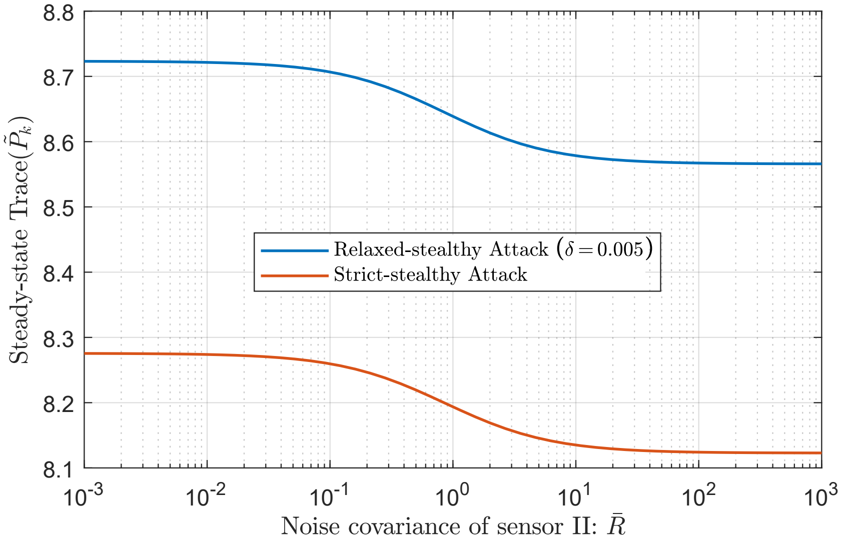

Figure 4.7: Impact of the side information on steady-state attack performance.

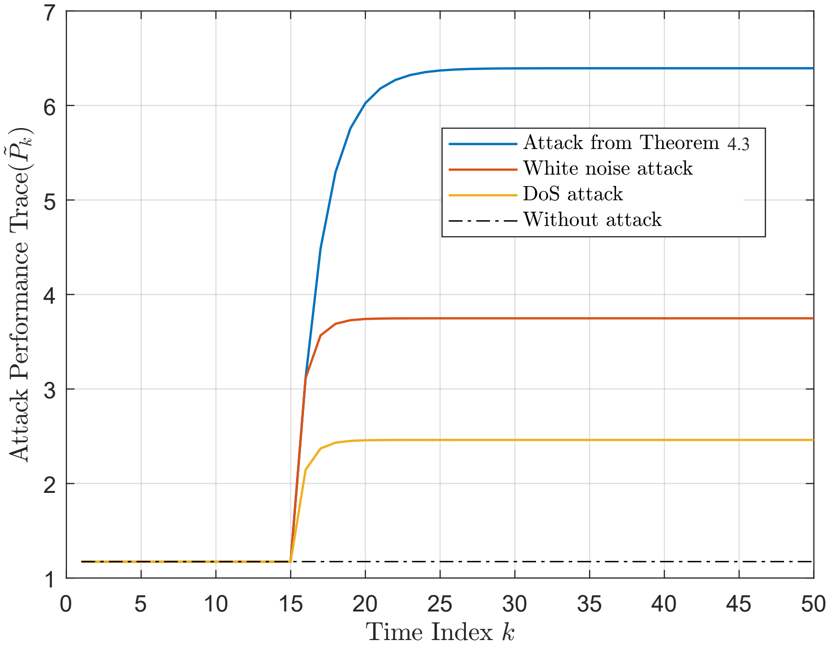

Figure 4.8: Attack performance with null information.

Chapter 5

Optimal Information-Based Deception Attacks with Multiple-Step Anomaly Detectors *

This chapter studies the problem of optimal deception attacks on remote state estimation, where an interval χ^2 detector is deployed to reveal anomalies. The information-based attack policy that can bypass the anomaly detector and cause the maximum estimation quality degradation is derived. For both attacks with strict and relaxed stealthiness, the optimal compromised measurements can be designed with three steps: obtain the minimum mean-square error estimation of the prediction error, de-correlate the estimate with historical compromised innovations, and design the compromised innovation as an optimal linear transformation. All available information for attackers is fully utilized for performance maximization while the stealthiness constraint is satisfied precisely to deceive the anomaly detector. The attack effect depends on both the amount of online information and the duration of detection interval. Contrary to well-studied innovation-based attacks using static linear

*A preliminary version of this chapter has been published as: Jing Zhou, Jun Shang, and Tongwen Chen, "Worst-case stealthy false-data injection attacks on remote state estimation," in *47th Annual Conference of the IEEE Industrial Electronics Society (IECON)*, Toronto, Oct. 2021. A full version entitled "Optimal deception attacks on remote state estimators equipped with interval anomaly detectors" has been submitted to *Automatica* for publication. (Provisionally accepted as a regular paper) combinations, the information-based deception policy is shown to be generated by a linear time-varying system, whose coefficients can be completely determined offline. The optimality of the proposed attack is verified with numerical examples and comparative studies.

This chapter is organized as follows. Section 5.1 describes the system model and formulates the deception attack problem. Section 5.2 focuses on optimal attacks with strict stealthiness. Section 5.3 studies optimal attacks with relaxed stealthiness. Section 5.4 uses numerical examples to verify the theoretical results. Finally, Section 5.5 concludes this paper.

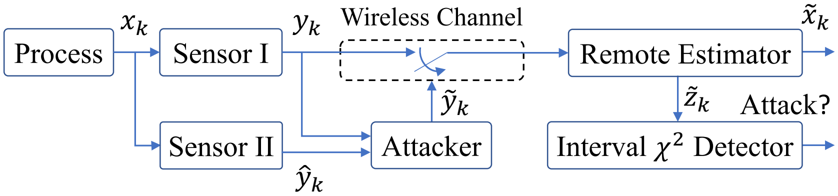

Figure 5.1: Deception attacks on remote state estimation with a multiple-step χ^2 detector.

5.1 Problem Formulation

The system architecture for remote state estimation and deception attacks is illustrated in Fig. 5.1.

5.1.1 Process Model

The discrete linear time-invariant process is given by

$$x_{k+1} = Ax_k + w_k, \tag{5.1}$$

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

where $k \in \mathbb{N}$ is the time index; $x_k \in \mathbb{R}^n$ denotes the system state, $y_k \in \mathbb{R}^m$ is the measurement of sensor I; $w_k \in \mathbb{R}^n$ and $v_k \in \mathbb{R}^m$ are i.i.d. Gaussian noises with covariance $Q \in \mathbb{S}_+^n$ and $R \in \mathbb{S}_{++}^m$, respectively. The initial state $x_0 \in \mathbb{R}^n$ is zero-mean Gaussian with covariance $\Pi_0 \in \mathbb{S}_+^n$, independent of w_k

and $v_k, \forall k \in \mathbb{N}$. Assume $m \leq n$, the pair $(A, \sqrt{Q})$ is stabilizable and (A, C) is detectable.

5.1.2 Remote Estimator

At nominal conditions, the measurement of sensor I is sent to the remote end sequentially through a wireless channel. A standard Kalman filter without packet loss and delays is adopted to estimate system states [3]:

$$\begin{aligned} x_{k|k-1} &= Ax_{k-1|k-1}, && (5.3a)\\ x_{k|k} &= x_{k|k-1} + K_k z_k, && (5.3b)\\ z_k &= y_k - Cx_{k|k-1}, && (5.3c)\\ K_k &= P_k^- C^{\mathrm{T}}(CP_k^- C^{\mathrm{T}} + R)^{-1}, && (5.3d)\\ P_{k|k-1} &= AP_{k-1|k-1}A^{\mathrm{T}} + Q, && (5.3e)\\ P_{k|k} &= (I_n - K_k C)P_{k|k-1}, && (5.3f) \end{aligned}$$

where $x_{k|k-1}$ and $x_{k|k}$ denote the *a priori* and *a posteriori* minimum mean-square error (MMSE) state estimates, respectively. $P_{k|k-1}$ and $P_{k|k}$ are the corresponding estimation error covariances. It is known that the Kalman filter converges from any initial condition. In steady state, $P_{k|k-1}$ is given by the unique solution of the Riccati equation:

$$\bar{P} = h[g_{[C,R]}(\bar{P})], \tag{5.4}$$

and the nominal innovation $z_k \in \mathbb{R}^m$ is zero-mean i.i.d. Gaussian with covariance $\Sigma = C\bar{P}C^{\mathrm{T}} + R$.

5.1.3 Anomaly Detector

To reveal potential component faults, transmission errors and cyber-attacks, we assume that a widely-used interval χ^2 detector is deployed at the remote end [22–24, 36]. Let $\tau \in \mathbb{N}$ denote the width of the sliding window (detection interval). At each sampling instant, the anomaly detector evaluates the following detection function

$$g_k(z_{k-\tau+1},\cdots,z_k)=\sum_{i=k-\tau+1}^{k} z_i^{\mathrm{T}}\Sigma^{-1}z_i, \tag{5.5}$$

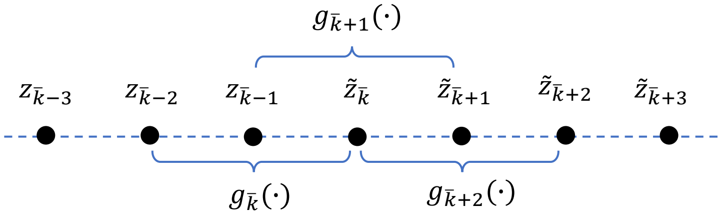

Figure 5.2: Interval χ^2 detector with $\tau = 3$.

5.1.4 Deception Attack

Since y_k is transmitted through an unreliable wireless link, a malicious opponent may implement a spurious transmitter to send falsified data to the receiver [22]. Assume the attack starts from instant $\bar{k}$. To investigate the impact of worst-case deception attacks, we assume that the attacker has the ability to modify y_k to $\tilde{y}_k$ and also

1. knows all system parameters and $x_{\bar{k}|\bar{k}-1}$,

2. can eavesdrop on the measurements of sensor I,

3. can also place an extra sensor (denoted as sensor II in Fig. 5.1) to obtain some side information of system states:

$$\hat{y}_k = \hat{C}x_k + \hat{v}_k,$$

   where $\hat{v}_k$ is i.i.d. zero-mean Gaussian with covariance $\hat{R} \in \mathbb{S}^{\bar{m}}_{++}$. Owing to common environmental disturbances, the measurement noises of the two sensors can be correlated, i.e., $\mathbb{E}[v_i \hat{v}_j^{\mathrm{T}}] = \delta_{ij} S$ with $S \in \mathbb{R}^{m \times \bar{m}}$.

At the kth sampling instant, the information available to the attacker is denoted by the set

$$\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k, \hat{y}_k\},\ \mathbb{I}_{\bar{k}-1} = \emptyset.$$

Note that $\mathbb{I}_k$ defines a general information set that covers different attack scenarios. If owing to some defensive countermeasures, no online measurements can be obtained, then $\mathbb{I}_k = \emptyset$. If only the original measurement can be intercepted, then $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k\}$. If only the side information is available, then $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{\hat{y}_k\}$. All these cases will be studied in a unified framework in this paper. Additionally, the assumption for knowing $x_{\bar{k}|\bar{k}-1}$ is standard in existing work on innovation-based linear attacks if the transmitted data is the raw measurement, i.e., sensor I is *not* a smart sensor [23, 24]. In this case, $x_{\bar{k}|\bar{k}-1}$ is indispensable to infer the nominal innovation.

Based on $\mathbb{I}_k$, the attacker sends the faked measurement $\tilde{y}_k$ to the remote end to degrade the estimation quality. Let $\tilde{x}_{k|k-1}$ and $\tilde{x}_{k|k}$ denote the *a priori* and *a posteriori* state estimates under deception attacks, respectively. $\tilde{P}_{k|k-1}$ and $\tilde{P}_{k|k}$ are the corresponding error covariances:

$$\tilde{P}_{k|k-1} = \mathbb{E}[(x_k - \tilde{x}_{k|k-1})(x_k - \tilde{x}_{k|k-1})^{\mathrm{T}}], \tag{5.7a}$$

$$\tilde{P}_{k|k} = \mathbb{E}[(x_k - \tilde{x}_{k|k})(x_k - \tilde{x}_{k|k})^{\mathrm{T}}]. \tag{5.7b}$$

The attack performance is evaluated by $\mathrm{Tr}(\tilde{P}_{k|k})$.

5.1.5 Problem of Interest

Owing to the existence of an interval χ^2 detector, the attacker should design the falsified data $\tilde{y}_k$ precisely to remain stealthy. In this paper, we study the optimal information-based attack policy

$$\tilde{y}_k = \pi_k(\mathbb{I}_k), \tag{5.8}$$

which can maximize $\mathrm{Tr}(\tilde{P}_{k|k})$ step by step while satisfying the following stealthiness constraints:

$$\mathcal{D}_{\mathrm{KL}}(\tilde{z}_k \| z_k) \leq \delta, \tag{5.9a}$$

$$\mathbb{E}[\tilde{z}_k \tilde{z}_i^{\mathrm{T}}] = 0_m, \ \forall i \in [\![k-\tau+1, k-1]\!], \tag{5.9b}$$

where $\tilde{z}_k$ is the compromised innovation, $\pi_k(\cdot)$ is a general attack strategy based on all available information, $\mathcal{D}_{\mathrm{KL}}(\cdot)$ denotes the KL divergence†, and $\delta \in \mathbb{R}_+$ is the stealthiness level. Since $\tilde{z}_k = \tilde{y}_k - C\tilde{x}_{k|k-1}$ where $\tilde{x}_{k|k-1}$ will be given by (5.12), designing $\tilde{z}_k$ is equivalent to designing $\tilde{y}_k$ [23]. Hereinafter, we use $\tilde{y}_k = \pi_k(\mathbb{I}_k)$ and $\tilde{z}_k = \pi_k(\mathbb{I}_k)$ interchangeably to represent the general attack strategy. Without loss of generality, we assume the remote estimator has entered steady state before $\bar{k}-\tau+1$. To ensure that (5.9b) is a consistent constraint in the early stage of deception attacks (see Fig. 5.2), we define $\tilde{z}_k = z_k, \forall k \in [\![\bar{k}-\tau+1, \bar{k}-1]\!]$. For clarity, (5.9b) is also called the "interval uncorrelation" constraint.

Remark 5.1. *In* (5.8), $\pi_k(\cdot)$ *denotes an information-based attack policy. Contrary to existing studies on deception attacks on remote state estimation [22–24, 36, 39, 67, 68, 72, 97], we do not presuppose that the attack is innovation based or has a linear form.*

Remark 5.2. *In existing studies, constraint* (5.9b) *has not been explicitly addressed. Specifically, in [22–24], and [97],* $\tilde{z}_k$ *was a linear function of the (current) nominal innovation, leading to i.i.d. compromised innovations. This attack could successfully bypass an interval* χ^2 *detector, but it posed an overly restrictive constraint on* $\tilde{z}_k$ *(*$\tau = \infty$*). In [68], an interval of historical nominal*

$\dagger$The KL divergence is a well-established metric that measures the statistical distance of two random variables. $\mathcal{D}_{\mathrm{KL}}(\tilde{z}_k \| z_k)$ can reflect the influence of deception attacks on the probability distribution of z_k, and thus is a reasonable choice to indicate the stealthiness level.

innovations was utilized to design a linear attack, but the resulting compromised innovations were correlated in every two consecutive steps. The corresponding attack could deceive only a single-step χ^2 detector ($\tau = 1$).

5.2 Optimal Attacks with Strict Stealthiness

In this section, we let $\delta = 0$ and derive the optimal deception attacks satisfying (5.9). It is known that $\tilde{z}_k$ and z_k have the same probability distribution with zero KL divergence; thus the alarm rate under deception attacks is also α. In this sense, the attack is called strictly stealthy. When $k \geq \bar{k}$, the estimator becomes:

$$\tilde{x}_{k|k-1} = A\tilde{x}_{k-1|k-1}, \tag{5.10}$$

$$\tilde{x}_{k|k} = \tilde{x}_{k|k-1} + K\tilde{z}_k, \tag{5.11}$$

where the fixed gain is $K = \bar{P}C^{\mathrm{T}}\Sigma^{-1}$. It follows that

$$\tilde{x}_{k|k-1} = A\tilde{x}_{k-1|k-2} + AK\tilde{z}_{k-1}. \tag{5.12}$$

Since the estimator is not altered before $\bar{k}$, the initial condition is $\tilde{x}_{\bar{k}|\bar{k}-1} = x_{\bar{k}|\bar{k}-1}$. According to (5.12), $\tilde{x}_{k|k-1}$ is determined by all compromised innovations in the interval $[\![\bar{k}, k-1]\!]$ and $x_{\bar{k}|\bar{k}-1}$, thus it is a known variable at the kth instant. From (5.10)–(5.11), (5.7) becomes

$$\tilde{P}_{k|k-1} = A\tilde{P}_{k-1|k-1}A^{\mathrm{T}} + Q, \tag{5.13a}$$

$$\tilde{P}_{k|k} = \tilde{P}_{k|k-1} + K\mathbb{E}[\tilde{z}_k\tilde{z}_k^{\mathrm{T}}]K^{\mathrm{T}} - K\mathbb{E}[\tilde{z}_k\tilde{e}_{k|k-1}^{\mathrm{T}}] - \mathbb{E}[\tilde{e}_{k|k-1}\tilde{z}_k^{\mathrm{T}}]K^{\mathrm{T}}, \tag{5.13b}$$

where $\tilde{e}_{k|k-1} = x_k - \tilde{x}_{k|k-1}$ denotes the prediction error with deception attacks. Note that $\tilde{P}_{k-1|k-1}$ is a constant at the kth instant. According to (5.13b), the design of $\tilde{z}_k$ can be formulated as

$$\begin{aligned} \mathbf{P}_{5.1}: \quad & \min_{\tilde{z}_k = \pi_k(\mathbb{I}_k)} \quad \mathrm{Tr}\{K\mathbb{E}[\tilde{z}_k\tilde{e}_{k|k-1}^{\mathrm{T}}]\} \\ & \qquad \text{s.t.} \ \ \tilde{z}_k \sim \mathcal{N}(0_{m\times 1}, \Sigma) \ \text{ and } \ (5.9\text{b}). \end{aligned}$$

Next we consider a special type of deception attacks and find the optimal solution to $\mathbf{P}_{5.1}$.

5.2.1 MMSE Estimation of Prediction Errors

Define the variables and matrices:

$$\bar{y}_k = \begin{bmatrix} y_k \\ \hat{y}_k \end{bmatrix}, \bar{v}_k = \begin{bmatrix} v_k \\ \hat{v}_k \end{bmatrix}, \bar{C} = \begin{bmatrix} C \\ \hat{C} \end{bmatrix}, \bar{R} = \begin{bmatrix} R & S \\ S^{\mathrm{T}} & \hat{R} \end{bmatrix}.$$

The online measurement becomes $\bar{y}_k = \bar{C}x_k + \bar{v}_k$. For plant (5.1), the attacker uses the following Kalman filter to obtain an optimal state estimation:

$$\bar{\alpha}_k = A\alpha_{k-1}, \tag{5.14a}$$

$$\alpha_k = \bar{\alpha}_k + \bar{K}_k(\bar{y}_k - \bar{C}\bar{\alpha}_k), \tag{5.14b}$$

$$\bar{K}_k = \bar{P}_k^e \bar{C}^{\mathrm{T}}(\bar{C}\bar{P}_k^e\bar{C}^{\mathrm{T}} + \bar{R})^{-1}, \tag{5.14c}$$

$$\bar{P}_k^e = AP_{k-1}^e A^{\mathrm{T}} + Q, \tag{5.14d}$$

$$P_k^e = (I_n - \bar{K}_k\bar{C})\bar{P}_k^e, \tag{5.14e}$$

where $\bar{\alpha}_k$ and α_k denote the *a priori* and *a posteriori* MMSE state estimates, respectively. $\bar{P}_k^e$ and P_k^e are the corresponding error covariances. Since the attacker does not have any online information but knows $x_{\bar{k}|\bar{k}-1}$ before instant $\bar{k}$, the initial state of the above Kalman filter is $\bar{\alpha}_{\bar{k}} = x_{\bar{k}|\bar{k}-1}, \bar{P}_{\bar{k}}^e = \bar{P}$. Recall that $\tilde{e}_{k|k-1} = x_k - \tilde{x}_{k|k-1}$ and $\tilde{x}_{k|k-1}$ is a constant at the kth instant; the MMSE estimate for $\tilde{e}_{k|k-1}$ is therefore given by

$$\beta_k = \mathbb{E}[\tilde{e}_{k|k-1}|\mathbb{I}_k] = \alpha_k - \tilde{x}_{k|k-1}. \tag{5.15}$$

Denote $P_k^\beta = \mathbb{E}[\beta_k\beta_k^{\mathrm{T}}]$. Since the MMSE estimate is orthogonal to the estimation error [3], we have

$$\begin{aligned} P_k^\beta &= \mathbb{E}[\tilde{e}_{k|k-1}\tilde{e}_{k|k-1}^{\mathrm{T}}] - \mathbb{E}[(\tilde{e}_{k|k-1} - \beta_k)(\tilde{e}_{k|k-1} - \beta_k)^{\mathrm{T}}] \\ &= \mathbb{E}[\tilde{e}_{k|k-1}\tilde{e}_{k|k-1}^{\mathrm{T}}] - \mathbb{E}[(x_k - \alpha_k)(x_k - \alpha_k)^{\mathrm{T}}] \\ &= \tilde{P}_{k|k-1} - P_k^e. \end{aligned} \tag{5.16}$$

Denote $\hat{e}_k = K^{\mathrm{T}}\tilde{e}_{k|k-1} \in \mathbb{R}^m$; the MMSE estimate of $\hat{e}_k$ and its covariance are given by

$$\phi_k = \mathbb{E}[\hat{e}_k|\mathbb{I}_k] = K^{\mathrm{T}}\beta_k, \tag{5.17}$$

$$P_k^\phi = \mathbb{E}[\phi_k\phi_k^{\mathrm{T}}] = K^{\mathrm{T}}P_k^\beta K. \tag{5.18}$$

The objective function of $\mathbf{P}_{5.1}$ becomes

$$\mathrm{Tr}\{\mathbb{E}[\tilde{z}_k(K^{\mathrm{T}}\tilde{e}_{k|k-1})^{\mathrm{T}}]\} = \mathrm{Tr}\{\mathbb{E}[\hat{e}_k\tilde{z}_k^{\mathrm{T}}]\}. \tag{5.19}$$

5.2.2 Linear Attack Based on MMSE Estimates

Let $\tau_k = \min\{\tau, k - \bar{k} + 1\}$, $\bar{\tau}_k = m\tau_k$ and define the following vector

$$\theta_k = \begin{bmatrix}\phi_k^{\mathrm{T}}, & \tilde{z}_{k-1}^{\mathrm{T}}, & \cdots, & \tilde{z}_{k-\tau_k+1}^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{\bar{\tau}_k}. \tag{5.20}$$

The deception attack based on the MMSE estimate of $\hat{e}_k$ is designed as

$$\tilde{z}_k = H_k\theta_k + b_k,\ b_k \sim \mathcal{N}(0_{m\times 1}, \Phi_k), \tag{5.21}$$

where b_k is an i.i.d. Gaussian sequence and independent of all other variables; $H_k \in \mathbb{R}^{m\times\bar{\tau}_k}$ and $\Phi_k \in \mathbb{S}_+^m$ are decision variables. We now give the following proposition to simplify the "interval uncorrelation" constraint.

Proposition 5.1. *The attack in* (5.21) *satisfies* $\mathbb{E}[\tilde{z}_k z_i^{\mathrm{T}}] = 0_m,\ \forall i < \bar{k} \leq k.$

Proof. See Appendix A.6. ■

Since $\tilde{z}_k$ is independent of z_i, in the early stage of deception attacks ($k < \bar{k} + \tau$), we need only to ensure that $\tilde{z}_k$ is uncorrelated with all compromised innovations in the interval $[\![\bar{k}, k-1]\!]$. The "interval uncorrelation" constraint in (5.9b) becomes

$$\mathbb{E}[\tilde{z}_k\tilde{z}_i^{\mathrm{T}}] = 0_m,\ \forall i \in [\![k - \tau_k + 1, k-1]\!]. \tag{5.22}$$

We now evaluate the objective function in (5.19):

$$\begin{aligned}
&\mathrm{Tr}\{\mathbb{E}[\hat{e}_k\tilde{z}_k^{\mathrm{T}}]\} = \mathrm{Tr}\{\mathbb{E}[\hat{e}_k\theta_k^{\mathrm{T}}]H_k^{\mathrm{T}}\}\\
&= \mathrm{Tr}\left\{\begin{bmatrix}\mathbb{E}[\hat{e}_k\phi_k^{\mathrm{T}}] & \mathbb{E}[\hat{e}_k\tilde{z}_{k-1}^{\mathrm{T}}] & \cdots & \mathbb{E}[\hat{e}_k\tilde{z}_{k-\tau_k+1}^{\mathrm{T}}]\end{bmatrix} H_k^{\mathrm{T}}\right\}\\
&= \mathrm{Tr}\left\{\begin{bmatrix}P_k^{\phi} & \mathcal{M}_{k,k-1} & \cdots & \mathcal{M}_{k,k-\tau_k+1}\end{bmatrix} H_k^{\mathrm{T}}\right\},
\end{aligned} \tag{5.23}$$

where $\mathbb{E}[\hat{e}_k\phi_k^{\mathrm{T}}] = \mathbb{E}[\phi_k\phi_k^{\mathrm{T}}] = P_k^{\phi}$, and $\mathcal{M}_{i,j}$ is defined as

$$\mathcal{M}_{i,j} = \mathbb{E}[\hat{e}_i\tilde{z}_j^{\mathrm{T}}] = \mathbb{E}[\phi_i\tilde{z}_j^{\mathrm{T}}],\ i,j \in \mathbb{N}, i-1 \geq j \geq \bar{k}.$$

The first stealthiness constraint in $\mathbf{P}_{5.1}$ becomes

$$H_k \begin{bmatrix} P_k^{\phi} & \mathcal{M}_{k,k-1} & \cdots & \mathcal{M}_{k,k-\tau_k+1} \\ \mathcal{M}_{k,k-1}^{\mathrm{T}} & \Sigma & \cdots & 0_m \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{M}_{k,k-\tau_k+1}^{\mathrm{T}} & 0_m & \cdots & \Sigma \end{bmatrix} H_k^{\mathrm{T}} + \Phi_k = \Sigma. \tag{5.24}$$

The first "interval uncorrelation" constraint in (5.22) is

$$\mathbb{E}[\tilde{z}_k \tilde{z}_{k-1}^{\mathrm{T}}] = H_k \mathbb{E}[\theta_k \tilde{z}_{k-1}^{\mathrm{T}}] = H_k \begin{bmatrix} \mathbb{E}[\phi_k \tilde{z}_{k-1}^{\mathrm{T}}] \\ \mathbb{E}[\tilde{z}_{k-1} \tilde{z}_{k-1}^{\mathrm{T}}] \\ \vdots \\ \mathbb{E}[\tilde{z}_{k-\tau_k+1} \tilde{z}_{k-1}^{\mathrm{T}}] \end{bmatrix} = H_k \begin{bmatrix} \mathcal{M}_{k,k-1} \\ \Sigma \\ \vdots \\ 0_m \end{bmatrix} = 0_m,$$

where we use the fact that $\tilde{z}_{k-i}$ and $\tilde{z}_{k-j}$ are uncorrelated $\forall i,j \in [\![1, \tau_k - 1]\!], i \neq j$. Similarly, all "interval uncorrelation" constraints can be written in a compact form as

$$H_k \begin{bmatrix} \mathcal{M}_{k,k-1} & \cdots & \mathcal{M}_{k,k-\tau_k+1} \\ \Sigma & \cdots & 0_m \\ \vdots & \ddots & \vdots \\ 0_m & \cdots & \Sigma \end{bmatrix} = 0_{m \times (\bar{\tau}_k - m)}. \tag{5.25}$$

Define the following matrices

$$\mathcal{M}_k = \begin{bmatrix} \mathcal{M}_{k,k-1} & \cdots & \mathcal{M}_{k,k-\tau_k+1} \end{bmatrix}, \tag{5.26}$$

$$\Sigma_k = \operatorname{blkdiag} \underbrace{\{\Sigma, \cdots, \Sigma\}}_{\tau_k - 1 \text{ times}}. \tag{5.27}$$

According to (5.23)–(5.25), the optimization problem becomes

$$\begin{aligned} \mathbf{P}_{5.2}: \quad & \min_{H_k \in \mathbb{R}^{m \times \bar{\tau}_k}, \Phi_k \in \mathbb{S}_+^m} \quad \mathrm{Tr}\left\{ \begin{bmatrix} P_k^{\phi} & \mathcal{M}_k \end{bmatrix} H_k^{\mathrm{T}} \right\} \\ & \text{s.t.} \quad H_k \begin{bmatrix} P_k^{\phi} & \mathcal{M}_k \\ \mathcal{M}_k^{\mathrm{T}} & \Sigma_k \end{bmatrix} H_k^{\mathrm{T}} + \Phi_k = \Sigma, \\ & \qquad H_k \begin{bmatrix} \mathcal{M}_k \\ \Sigma_k \end{bmatrix} = 0_{m \times (\bar{\tau}_k - m)}. \end{aligned}$$

Remark 5.3. *Without knowing the structure of $\tilde{z}_k$, the original problem $\boldsymbol{P}_{5.1}$ is not a standard optimization problem. By assuming a linear attack model based on the MMSE estimate, it is reformulated as a convex optimization problem. In next section we show how to obtain the constant parameters in $\boldsymbol{P}_{5.2}$.*

5.2.3 Recursion of Parameters

The recursions of P_k^ϕ and Σ_k are given by (5.18) and (5.27), respectively. From (5.14a)–(5.14b), we have

$$\alpha_k = A\alpha_{k-1} - \bar{K}_k\bar{C}\bar{\alpha}_k + \bar{K}_k\bar{y}_k. \tag{5.28}$$

Combining (5.12), (5.28) and $\bar{y}_k = \bar{C}x_k + \bar{v}_k$ yields

$$\beta_k = A\beta_{k-1} + \bar{K}_k\bar{C}(x_k - \bar{\alpha}_k) + \bar{K}_k\bar{v}_k - AK\tilde{z}_{k-1}. \tag{5.29}$$

Since $\tilde{z}_{k-1}$ is based on the information set $\mathbb{I}_{k-1}$ while $x_k - \bar{\alpha}_k$ is independent of $\mathbb{I}_{k-1}$ (by the orthogonality property of MMSE estimation), and $\bar{v}_k$ is independent of $\tilde{z}_{k-1}$, (5.29) implies that

$$\mathbb{E}[\beta_k\tilde{z}_{k-1}^{\mathrm{T}}] = A\mathbb{E}[\beta_{k-1}\tilde{z}_{k-1}^{\mathrm{T}}] - AK\mathbb{E}[\tilde{z}_{k-1}\tilde{z}_{k-1}^{\mathrm{T}}]. \tag{5.30}$$

We now evaluate the second term in (5.30). Because b_k is independent of β_k, we have

$$\begin{aligned}\mathbb{E}[\beta_k\tilde{z}_k^{\mathrm{T}}] &= \mathbb{E}[\beta_k(H_k\theta_k + b_k)^{\mathrm{T}}] = \mathbb{E}[\beta_k\theta_k^{\mathrm{T}}]H_k^{\mathrm{T}} \\ &= \begin{bmatrix}\mathbb{E}[\beta_k\phi_k^{\mathrm{T}}] & \mathbb{E}[\beta_k\tilde{z}_{k-1}^{\mathrm{T}}] & \cdots & \mathbb{E}[\beta_k\tilde{z}_{k-\tau_k+1}^{\mathrm{T}}]\end{bmatrix} H_k^{\mathrm{T}}.\end{aligned} \tag{5.31}$$

Define the matrix

$$\mathcal{W}_{i,j} = \mathbb{E}[\beta_i\tilde{z}_j^{\mathrm{T}}],\ i,j \in \mathbb{N}, i-1 \geq j \geq \bar{k}. \tag{5.32}$$

Since $\mathbb{E}[\beta_k\phi_k^{\mathrm{T}}] = \mathbb{E}[\beta_k\beta_k^{\mathrm{T}}]K = P_k^\beta K$, it follows from (5.30)–(5.31) that

$$\mathcal{W}_{k,k-1} = A\begin{bmatrix}P_{k-1}^\beta K & \mathcal{W}_{k-1,k-2} & \cdots & \mathcal{W}_{k-1,k-\tau_{k-1}}\end{bmatrix} H_{k-1}^{\mathrm{T}} - AK\Sigma. \tag{5.33}$$

Similarly, when $i \geq 2$, from (5.29) we have

$$\mathbb{E}[\beta_k\tilde{z}_{k-i}^{\mathrm{T}}] = A\mathbb{E}[\beta_{k-1}\tilde{z}_{k-i}^{\mathrm{T}}],$$

where again we use the fact that $\tilde{z}_{k-i}$ is independent of $\tilde{z}_{k-1}$, $\bar{v}_k$ and $x_k - \bar{\alpha}_k$. It follows that

$$\mathcal{W}_{k,k-i} = A\mathcal{W}_{k-1,k-i},\ \forall i \in [\![2, \tau_k - 1]\!]. \tag{5.34}$$

Note that $\mathcal{M}_{i,j} = K^{\mathrm{T}}\mathbb{E}[\beta_i \tilde{z}_j^{\mathrm{T}}] = K^{\mathrm{T}}\mathcal{W}_{i,j}$. The parameter $\mathcal{M}_k$ in $\mathbf{P}_{5.2}$ is recursively determined by (5.33) and (5.34). At the $\bar{k}$th sampling instant, $\theta_{\bar{k}} = \phi_{\bar{k}}$, then $\mathcal{M}_{\bar{k}}$ and $\Sigma_{\bar{k}}$ vanish. The optimal coefficient $H_{\bar{k}}^*$ can be obtained by solving $\mathbf{P}_{5.2}$. When $k = \bar{k}+1$, we have

$$\mathcal{W}_{\bar{k}+1,\bar{k}} = AP_{\bar{k}}^{\beta}KH_{\bar{k}}^* - AK\Sigma. \tag{5.35}$$

This is the initial condition of parameter recursion.

5.2.4 Explicit Solution of Attack Coefficients

In this section we derive the explicit optimal solution to $\mathbf{P}_{5.2}$. The main result is summarized in the following lemma.

Lemma 5.1. *The optimal solution to $\boldsymbol{P}_{5.2}$ is given by*

$$H_k^* = -\Sigma^{\frac{1}{2}}(U_k S_k^{-1} U_k^{\mathrm{T}} + \mathcal{Z}_k \tilde{U}_k^{\mathrm{T}})\Sigma^{\frac{1}{2}} \begin{bmatrix} I_m & -\mathcal{M}_k\Sigma_k^{-1} \end{bmatrix},$$

$$\Phi_k^* = \Sigma^{\frac{1}{2}}\tilde{U}_k\tilde{U}_k^{\mathrm{T}}\Sigma^{\frac{1}{2}},$$

where U_k and S_k satisfy the economy-size SVD:

$$\Sigma^{\frac{1}{2}}(P_k^{\phi} - \mathcal{M}_k\Sigma_k^{-1}\mathcal{M}_k^{\mathrm{T}})^{\frac{1}{2}} = U_k S_k V_k^{\mathrm{T}},\ S_k \in \mathbb{S}_{++}^{r_k};$$

$\tilde{U}_k$ is the orthogonal complement of U_k. $P_k^{\phi}, \mathcal{M}_k$ and Σ_k are recursively given by (5.18), (5.26) *and* (5.27), *respectively. If $\tau_k = 1$, then $\theta_k = \phi_k$, $\mathcal{M}_k$ and Σ_k vanish*‡. *$\mathcal{Z}_k \in \mathbb{R}^{m\times(m-r_k)}$ is a matrix of free entries.*

Proof. At the kth sampling instant, define $\Delta_k = \begin{bmatrix}\bar{\Delta}_k & \hat{\Delta}_k\end{bmatrix} \in \mathbb{R}^{m\times\bar{\tau}_k}$ such that

$$\Delta_k \begin{bmatrix}\mathcal{M}_k \\ \Sigma_k\end{bmatrix} = \bar{\Delta}_k\mathcal{M}_k + \hat{\Delta}_k\Sigma_k = 0_{m\times(\bar{\tau}_k - m)}, \tag{5.36}$$

then H_k can be parameterized as $H_k = \bar{H}_k\Delta_k$, where $\bar{H}_k \in \mathbb{R}^{m\times m}$ is a matrix of free entries. Note that the second constraint in $\mathbf{P}_{5.2}$ is eliminated. Since

‡This happens in two cases: *i*), $\forall\tau \in \mathbb{N}, \tau_{\bar{k}} = 1$. *ii*), if $\tau = 1$, then $\tau_k = 1$, $\forall k \geq \bar{k}$.

Σ_k is non-singular when $\tau_k \neq 1$, we have $\hat{\Delta}_k = -\bar{\Delta}_k \mathcal{M}_k \Sigma_k^{-1}$. The objective function becomes

$$\begin{aligned}
&\operatorname{Tr}\left\{\begin{bmatrix} P_k^{\phi} & \mathcal{M}_k \end{bmatrix} H_k^{\mathrm{T}}\right\} = \operatorname{Tr}\left\{\begin{bmatrix} P_k^{\phi} & \mathcal{M}_k \end{bmatrix} \begin{bmatrix} \bar{\Delta}_k^{\mathrm{T}} \\ \hat{\Delta}_k^{\mathrm{T}} \end{bmatrix} \bar{H}_k^{\mathrm{T}}\right\} \\
=&\operatorname{Tr}\left\{(P_k^{\phi}\bar{\Delta}_k^{\mathrm{T}} + \mathcal{M}_k\hat{\Delta}_k^{\mathrm{T}})\bar{H}_k^{\mathrm{T}}\right\} = \operatorname{Tr}\left\{(P_k^{\phi} - \mathcal{M}_k\Sigma_k^{-1}\mathcal{M}_k^{\mathrm{T}})\bar{\Delta}_k^{\mathrm{T}}\bar{H}_k^{\mathrm{T}}\right\}. \qquad (5.37)
\end{aligned}$$

The stealthiness constraint is

$$\begin{aligned}
&H_k \begin{bmatrix} P_k^{\phi} & \mathcal{M}_k \\ \mathcal{M}_k^{\mathrm{T}} & \Sigma_k \end{bmatrix} H_k^{\mathrm{T}} + \Phi_k \\
=&\bar{H}_k \begin{bmatrix} \bar{\Delta}_k & \hat{\Delta}_k \end{bmatrix} \begin{bmatrix} P_k^{\phi} & \mathcal{M}_k \\ \mathcal{M}_k^{\mathrm{T}} & \Sigma_k \end{bmatrix} \begin{bmatrix} \bar{\Delta}_k^{\mathrm{T}} \\ \hat{\Delta}_k^{\mathrm{T}} \end{bmatrix} \bar{H}_k^{\mathrm{T}} + \Phi_k \\
=&\bar{H}_k\bar{\Delta}_k \begin{bmatrix} I_m & -\mathcal{M}_k\Sigma_k^{-1} \end{bmatrix} \begin{bmatrix} P_k^{\phi} & \mathcal{M}_k \\ \mathcal{M}_k^{\mathrm{T}} & \Sigma_k \end{bmatrix} \begin{bmatrix} I_m \\ -\Sigma_k^{-1}\mathcal{M}_k^{\mathrm{T}} \end{bmatrix} \bar{\Delta}_k^{\mathrm{T}}\bar{H}_k^{\mathrm{T}} + \Phi_k \\
=&\bar{H}_k\bar{\Delta}_k(P_k^{\phi} - \mathcal{M}_k\Sigma_k^{-1}\mathcal{M}_k^{\mathrm{T}})\bar{\Delta}_k^{\mathrm{T}}\bar{H}_k^{\mathrm{T}} + \Phi_k = \Sigma. \qquad (5.38)
\end{aligned}$$

Define the temporary variables

$$\begin{aligned}
\hat{H}_k &= \Sigma^{-\frac{1}{2}}\bar{H}_k\bar{\Delta}_k(P_k^{\phi} - \mathcal{M}_k\Sigma_k^{-1}\mathcal{M}_k^{\mathrm{T}})^{\frac{1}{2}}, \\
\hat{Y}_k &= \Sigma^{\frac{1}{2}}(P_k^{\phi} - \mathcal{M}_k\Sigma_k^{-1}\mathcal{M}_k^{\mathrm{T}})^{\frac{1}{2}},\ r_k = \operatorname{rank}(\hat{Y}_k), \\
\bar{\Phi}_k &= \Sigma^{-\frac{1}{2}}\Phi_k\Sigma^{-\frac{1}{2}}.
\end{aligned}$$

According to (5.37)–(5.38), $\mathbf{P}_{5.2}$ becomes

$$\min_{\hat{H}_k \in \mathbb{R}^{m\times m}, \bar{\Phi}_k \in \mathbb{S}_+^m} \quad \operatorname{Tr}(\hat{Y}_k\hat{H}_k^{\mathrm{T}}) \qquad (5.39)$$

$$\text{s.t.} \quad \hat{H}_k\hat{H}_k^{\mathrm{T}} + \bar{\Phi}_k = I_m. \qquad (5.40)$$

Now perform economy-size singular value decomposition (SVD): $\hat{Y}_k = U_k S_k V_k^{\mathrm{T}}$. Subsequently, the objective function in (5.39) satisfies

$$\operatorname{Tr}(\hat{Y}_k\hat{H}_k^{\mathrm{T}}) = \operatorname{Tr}(U_kS_kV_k^{\mathrm{T}}\hat{H}_k^{\mathrm{T}}) = \operatorname{Tr}(S_kV_k^{\mathrm{T}}\hat{H}_k^{\mathrm{T}}U_k) = \sum_{i=1}^{r_k} S_k^{[i,i]}\tilde{H}_k^{[i,i]}, \qquad (5.41)$$

where $\tilde{H}_k = V_k^{\mathrm{T}}\hat{H}_k^{\mathrm{T}}U_k \in \mathbb{R}^{r_k\times r_k}$ and $X^{[i,j]}$ denotes the (i,j)th entry of X. It is clear that $\tilde{H}_k\tilde{H}_k^{\mathrm{T}} \preceq I_{r_k}$, leading to $H_k^{[i,i]} \in [-1,1], \forall i \in [\![1, r_k]\!]$. It follows from (5.41) that

$$\operatorname{Tr}(\hat{Y}_k\hat{H}_k^{\mathrm{T}}) \geq -\sum_{i=1}^{r_k} S_k^{[i,i]} = -\operatorname{Tr}(S_k), \qquad (5.42)$$

where the equality is attained only when $\tilde{H}_k^* = V_k^{\mathrm{T}} \hat{H}_k^{\mathrm{T}} U_k = -I_{r_k}$. Solving this matrix equation, we have

$$\hat{H}_k^* = -U_k V_k^{\mathrm{T}} + \tilde{U}_k \mathcal{X}_k + \mathcal{Y}_k \tilde{V}_k^{\mathrm{T}}, \tag{5.43}$$

where $\mathcal{X}_k \in \mathbb{R}^{(m-r_k)\times m}$ and $\mathcal{Y}_k \in \mathbb{R}^{m\times(m-r_k)}$ are matrices of free entries; $\tilde{U}_k$ and $\tilde{V}_k$ are orthogonal complements of U_k and V_k, respectively. Note that we have obtained the optimal solution to (5.39). The next step is to derive H_k^* from $\hat{H}_k^*$. According to the definitions of $\hat{H}_k$ and $\hat{Y}_k$, we have

$$\hat{Y}_k^{\mathrm{T}} \Sigma^{-\frac{1}{2}} (\bar{H}_k^* \bar{\Delta}_k)^{\mathrm{T}} \Sigma^{-\frac{1}{2}} = (\hat{H}_k^*)^{\mathrm{T}}.$$

The matrix equation yields

$$\Sigma^{-\frac{1}{2}} (\bar{H}_k^* \bar{\Delta}_k)^{\mathrm{T}} \Sigma^{-\frac{1}{2}} = (\hat{Y}_k^{\mathrm{T}})^{+} (\hat{H}_k^*)^{\mathrm{T}} + [I_m - (\hat{Y}_k^{\mathrm{T}})^{+} \hat{Y}_k^{\mathrm{T}}] \bar{\mathcal{Z}}_k,$$

where $\bar{\mathcal{Z}} \in \mathbb{R}^{m\times m}$ is a free parameter. Substituting $(\hat{Y}_k^{\mathrm{T}})^{+} = U_k S_k^{-1} V_k^{\mathrm{T}}$ and (5.43) into the above equation, we obtain

$$\bar{H}_k^* \bar{\Delta}_k = -\Sigma^{\frac{1}{2}} (U_k S_k^{-1} U_k^{\mathrm{T}} - \tilde{U}_k \mathcal{X}_k V_k S_k^{-1} U_k^{\mathrm{T}} - \bar{\mathcal{Z}}_k^{\mathrm{T}} \tilde{U}_k \tilde{U}_k^{\mathrm{T}}) \Sigma^{\frac{1}{2}}. \tag{5.44}$$

Note that $\hat{Y}_k \hat{Y}_k^{\mathrm{T}} = U_k S_k^2 U_k^{\mathrm{T}}$. The stealthiness constraint in (5.38) becomes

$$\begin{aligned}
&\bar{H}_k^* \bar{\Delta}_k \Sigma^{-\frac{1}{2}} \hat{Y}_k \hat{Y}_k^{\mathrm{T}} \Sigma^{-\frac{1}{2}} (\bar{H}_k^* \bar{\Delta}_k)^{\mathrm{T}} + \Phi_k^* \\
=&(\bar{H}_k^* \bar{\Delta}_k \Sigma^{-\frac{1}{2}} U_k S_k)(\bar{H}_k^* \bar{\Delta}_k \Sigma^{-\frac{1}{2}} U_k S_k)^{\mathrm{T}} + \Phi_k^* \\
=&\Sigma^{\frac{1}{2}} (U_k - \tilde{U}_k \mathcal{X}_k V_k)(U_k - \tilde{U}_k \mathcal{X}_k V_k)^{\mathrm{T}} \Sigma^{\frac{1}{2}} + \Phi_k^* = \Sigma.
\end{aligned}$$

Since $\Phi_k^* \in \mathbb{S}_+^m$, the following inequality holds:

$$\begin{bmatrix} U_k & \tilde{U}_k \end{bmatrix} \begin{bmatrix} I_{r_k} & -V_k^{\mathrm{T}} \mathcal{X}_k^{\mathrm{T}} \\ -\mathcal{X}_k V_k & \mathcal{X}_k V_k V_k^{\mathrm{T}} \mathcal{X}_k^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} U_k^{\mathrm{T}} \\ \tilde{U}_k^{\mathrm{T}} \end{bmatrix} \preceq I_m, \tag{5.45}$$

which directly leads to $\mathcal{X}_k V_k = 0_{(m-r_k)\times r_k}$. Then (5.44) reduces to

$$\bar{H}_k^* \bar{\Delta}_k = -\Sigma^{\frac{1}{2}} (U_k S_k^{-1} U_k^{\mathrm{T}} + \mathcal{Z}_k \tilde{U}_k^{\mathrm{T}}) \Sigma^{\frac{1}{2}}, \tag{5.46}$$

where the free parameter is re-defined as $\mathcal{Z}_k = -\bar{\mathcal{Z}}_k^{\mathrm{T}} \tilde{U}_k$. From (5.38), the covariance of b_k is given by $\Phi_k^* = \Sigma^{\frac{1}{2}} \tilde{U}_k \tilde{U}_k^{\mathrm{T}} \Sigma^{\frac{1}{2}}$. Recall that

$$H_k^* = \bar{H}_k^* \Delta_k = \bar{H}_k^* \bar{\Delta}_k \begin{bmatrix} I_m & -\mathcal{M}_k \Sigma_k^{-1} \end{bmatrix}. \tag{5.47}$$

Substituting (5.46) into the above equation yields H_k^*. ■

Remark 5.4. *In* (5.36)*,* Δ_k *is defined to eliminate the "interval uncorrelation" constraint. From* (5.46)*–*(5.47)*, it is not necessary to explicitly calculate* Δ_k *at each sampling instant. The explicit solution of optimal attack coefficients not only reduces the computational burden, but also serves as a key ingredient in Theorem 5.1 to establish the optimality of the attack in* (5.21)*.*

Since P_k^{ϕ} in (5.18) depends explicitly on $\tilde{P}_{k|k-1}$, which by (5.13) is determined by $\tilde{P}_{k-1|k-1}$, to make the recursion a closed loop, $\tilde{P}_{k-1|k-1}$ should be evaluated at each sampling step. From the definition of $\hat{e}_k$, we have $\tilde{e}_{k|k-1} = (K^{\mathrm{T}})^{+}\hat{e}_k + [I_n - (K^{\mathrm{T}})^{+}K^{\mathrm{T}}]\epsilon_k$, where $\epsilon_k \in \mathbb{R}^n$ is an arbitrary vector. It follows that

$$\mathbb{E}[\tilde{e}_{k|k-1}\tilde{z}_k^{\mathrm{T}}]K^{\mathrm{T}} = (K^{\mathrm{T}})^{+}\mathbb{E}[\hat{e}_k\tilde{z}_k^{\mathrm{T}}]K^{\mathrm{T}} + [I_n - (K^{\mathrm{T}})^{+}K^{\mathrm{T}}]\mathbb{E}[\epsilon_k\tilde{z}_k^{\mathrm{T}}]K^{\mathrm{T}}. \quad (5.48)$$

The second term has zero trace, thus it does not affect the attack performance. For simplicity, one can choose $\epsilon_k = 0_{n\times 1}$, leading to $\mathbb{E}[\tilde{e}_{k|k-1}\tilde{z}_k^{\mathrm{T}}]K^{\mathrm{T}} = (K^{\mathrm{T}})^{+}\mathbb{E}[\hat{e}_k\tilde{z}_k^{\mathrm{T}}]K^{\mathrm{T}}$. Then (5.13b) becomes

$$\begin{aligned}\tilde{P}_{k|k} = \tilde{P}_{k|k-1} &+ K\Sigma K^{\mathrm{T}} - KH_k^{*}\begin{bmatrix}P_k^{\phi} & \mathcal{M}_k\end{bmatrix}^{\mathrm{T}}K^{+}\\ &-(K^{\mathrm{T}})^{+}\begin{bmatrix}P_k^{\phi} & \mathcal{M}_k\end{bmatrix}(H_k^{*})^{\mathrm{T}}K^{\mathrm{T}}. \quad (5.49)\end{aligned}$$

Remark 5.5. *We see from the objective function of* $\boldsymbol{P}_{5.1}$ *that if* $\mathrm{rank}(K) < m$*, the attacker can always design a part of* $\tilde{z}_k$ *that lies in* $\mathrm{Ker}(K)$*. This component does not affect the attack performance but will impact the stealthiness constraint. The definition of* $\hat{e}_k$ *eliminates this phenomenon; but it should be pointed out that the technique is valid only when we consider the "greedy" attack policy. That is, the attack only maximizes* $\mathrm{Tr}(\tilde{P}_{k|k})$ *provided that* $\tilde{P}_{k-1|k-1}$ *is given. In this setting we can design* $\epsilon_k = 0_{n\times 1}$ *because the choice of* ϵ_k *has no impact on* $\mathrm{Tr}(\tilde{P}_{k|k})$*. Otherwise, if an interval attack performance is considered [38], different values of* ϵ_k *lead to the same* $\mathrm{Tr}\{\mathbb{E}[\tilde{e}_{k|k-1}\tilde{z}_k^{\mathrm{T}}]K^{\mathrm{T}}\}$ *but different* $\mathbb{E}[\tilde{e}_{k|k-1}\tilde{z}_k^{\mathrm{T}}]K^{\mathrm{T}}$*, which, by the recursion in* (5.13)*, will affect the attack performance in subsequent steps; thus the influence of* $\mathrm{Ker}(K)$ *cannot be ignored.*

5.2.5 Optimal Attack Policy

In the above analysis, an optimal linear attack based on the MMSE estimate of $\hat{e}_k$ is derived. In this section, we prove that this attack strategy is indeed the optimal one among all feasible policies satisfying the stealthiness constraints. The following theorem summarizes the main results in this paper.

Theorem 5.1. *The optimal deception attack policy with strict stealthiness is given by*

$$\tilde{y}_k^* = C\tilde{x}_{k|k-1} + H_k^*\theta_k + b_k,\ b_k \sim \mathcal{N}(0_{m\times 1}, \Phi_k^*), \tag{5.50}$$

where $\tilde{x}_{k|k-1}$ and θ_k are given by (5.12) *and* (5.20)*, respectively, and H_k^* and Φ_k^* are given in Lemma 5.1.*

Proof. The optimal compromised innovation from Theorem 5.1 is $\tilde{z}_k^* = H_k^*\theta_k + b_k$. b_k is only a compensation noise to ensure stealthiness; thus it has no impact on the objective function of $\mathbf{P}_{5.1}$. Since Φ_k^* is a constant, one can verify that $\hat{z}_k^* = H_k^*\theta_k$ is the optimal linear attack based on MMSE estimate that is derived from the following optimization problem (similar to $\mathbf{P}_{5.1}$):

$$\begin{aligned}\mathbf{P}_{5.3}: \quad &\min_{\tilde{z}_k=\pi_k(\mathbb{I}_k)} \quad \mathrm{Tr}\{\mathbb{E}[\tilde{z}_k\hat{e}_k^{\mathrm{T}}]\}\\ &\qquad\quad \text{s.t. } \ \tilde{z}_k \sim \mathcal{N}(0_{m\times 1}, \Sigma - \Phi_k^*) \ \text{ and } \ (5.22).\end{aligned}$$

Assume $\hat{z}_k = \pi_k(\mathbb{I}_k)$ is an arbitrary attack policy satisfying the above constraints. Define the following matrices associated with the objective functions of $\hat{z}_k$ and $\hat{z}_k^*$:

$$\hat{\mathcal{F}}_k = \mathbb{E}[\hat{z}_k\hat{e}_k^{\mathrm{T}}],\ \hat{\mathcal{F}}_k^* = \mathbb{E}[\hat{z}_k^*\hat{e}_k^{\mathrm{T}}].$$

Let $\Pi_k = -\bar{H}_k\bar{\Delta}_k$. From (5.47), we have

$$\hat{z}_k^* = H_k^*\theta_k = -\Pi_k\begin{bmatrix} I_m & -\mathcal{M}_k\Sigma_k^{-1}\end{bmatrix}\theta_k = -\Pi_k\bar{\theta}_k, \tag{5.51}$$

where $\bar{\theta}_k \in \mathbb{R}^m$ is given by

$$\bar{\theta}_k = \phi_k - \mathcal{M}_k\Sigma_k^{-1}\begin{bmatrix}\tilde{z}_{k-1}^{\mathrm{T}}, & \cdots, & \tilde{z}_{k-\tau_k+1}^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}. \tag{5.52}$$

Define $\mathbb{I}_k^z = \{\tilde{z}_{k-1}, \cdots, \tilde{z}_{k-\tau_k+1}\}$. For brevity, the set that contains all information in $\mathbb{I}_k$ and excludes all information in $\mathbb{I}_k^z$ is denoted as $\mathbb{I}_k^{y\backslash z}$. Note that $\phi_k = \mathbb{E}[\hat{e}_k|\mathbb{I}_k]$. Since θ_k is jointly Gaussian, (5.52) can be written as:

$$\bar{\theta}_k = \mathbb{E}[\hat{e}_k|\mathbb{I}_k^{y\backslash z}]. \tag{5.53}$$

At the kth sampling instant, Π_k is a constant; then $\hat{z}_k^*$ is the MMSE estimate of $-\Pi_k\hat{e}_k$ conditioned on $\mathbb{I}_k^{y\backslash z}$. To satisfy the "interval uncorrelation" constraint, $\hat{z}_k$ must also be designed based on $\mathbb{I}_k^{y\backslash z}$. The following matrix inequality holds:

$$\mathbb{E}[(-\Pi_k\hat{e}_k - \hat{z}_k^*)(-\Pi_k\hat{e}_k - \hat{z}_k^*)^{\mathrm{T}}] \preceq \mathbb{E}[(-\Pi_k\hat{e}_k - \hat{z}_k)(-\Pi_k\hat{e}_k - \hat{z}_k)^{\mathrm{T}}]. \tag{5.54}$$

Expanding the inequality and canceling identical terms, we have

$$\Pi_k(\hat{\mathcal{F}}_k^* - \hat{\mathcal{F}}_k)^{\mathrm{T}} + (\hat{\mathcal{F}}_k^* - \hat{\mathcal{F}}_k)\Pi_k^{\mathrm{T}} \preceq 0. \tag{5.55}$$

In (5.46), $\mathcal{Z}_k$ is a parameter that can be designed freely. Let $\mathcal{Z}_k = \lambda\tilde{U}_k$, where $\lambda > 0$; then

$$\begin{aligned}\Pi_k &= \Sigma^{\frac{1}{2}}(U_k S_k^{-1} U_k^{\mathrm{T}} + \lambda\tilde{U}_k\tilde{U}_k^{\mathrm{T}})\Sigma^{\frac{1}{2}} \\ &= \Sigma^{\frac{1}{2}} \begin{bmatrix} U_k & \tilde{U}_k \end{bmatrix} \begin{bmatrix} S_k^{-1} & 0_{r_k\times(m-r_k)} \\ 0_{(m-r_k)\times r_k} & \lambda I_{m-r_k} \end{bmatrix} \begin{bmatrix} U_k^{\mathrm{T}} \\ \tilde{U}_k^{\mathrm{T}} \end{bmatrix} \Sigma^{\frac{1}{2}} \succ 0. \end{aligned} \tag{5.56}$$

By Lyapunov stability theory and (5.55)–(5.56), all eigenvalues of $\hat{\mathcal{F}}_k^* - \hat{\mathcal{F}}_k$ have non-positive real parts. It follows that

$$\mathrm{Tr}(\hat{\mathcal{F}}_k^* - \hat{\mathcal{F}}_k) = \sum_{i=1}^{m} \Re_{\lambda_i}(\hat{\mathcal{F}}_k^* - \hat{\mathcal{F}}_k) \le 0. \tag{5.57}$$

The inequality implies that the objective value of $\hat{z}_k^*$ for $\mathbf{P}_{5.3}$ is no larger than that of $\hat{z}_k$. Since $\hat{z}_k$ is an arbitrary attack policy by assumption, the above analysis proves the optimality of $\hat{z}_k^*$. ■

Remark 5.6. *There exist multiple optimal attacks owing to the freedom to design $\mathcal{Z}_k$. All of them have the same attack performance. If $P_k^{\phi} - \mathcal{M}_k\Sigma_k^{-1}\mathcal{M}_k^{\mathrm{T}}$ is non-singular, i.e., $r_k = m$, $\mathcal{Z}_k$ and $\tilde{U}_k$ vanish; then the optimal compromised innovation is uniquely given by $\tilde{z}_k^* = H_k^*\theta_k$. If $r_k < m$, the simplest design*

Algorithm 5.1 Design of optimal attacks in Theorem 5.1

```
**Input**: Online measurements $y_k, \hat{y}_k$
**Output**: Optimal compromised measurement $\tilde{y}_k^*$
Initialize $\bar{\alpha}_{\bar{k}} = \hat{x}_{\bar{k}}^-$, $\tilde{x}_{\bar{k}}^- = \hat{x}_{\bar{k}}^-$, $\bar{P}_{\bar{k}}^e = \bar{P}$, $\tilde{P}_{\bar{k}|\bar{k}-1} = \bar{P}$.
**for** $k = \bar{k} : \infty$ **do**
    Set $\tau_k = \min\{\tau, k - \bar{k} + 1\}$, $\bar{\tau}_k = m\tau_k$.
    Run the filter in (5.14) to obtain $\alpha_k, P_k^e$.
    Calculate $\phi_k, P_k^\phi$ with (5.17)–(5.18).
    **if** $\tau_k = 1$ **then**
        Set $\theta_k = \phi_k$, $\hat{Y}_k = \Sigma^{\frac{1}{2}}(P_k^\phi)^{\frac{1}{2}}$.
    **else**
        Set $\theta_k$ as (5.20), $\hat{Y}_k = \Sigma^{\frac{1}{2}}(P_k^\phi - \mathcal{M}_k \Sigma_k^{-1} \mathcal{M}_k^{\mathrm{T}})^{\frac{1}{2}}$.
    **end if**
    Do SVD: $\hat{Y}_k = U_k S_k V_k^{\mathrm{T}}$, design $\tilde{U}_k, \mathcal{Z}_k$.
    Design $H_k^*$, $\Phi_k^*$, $\tilde{z}_k^*$, $\tilde{y}_k^*$ with Theorem 5.1.
    Evaluate $\tilde{P}_k$ with (5.49).
    **if** $\tau_k \neq 1$ **then**
        Update $\mathcal{W}_{k+1,k}$ with (5.33).
        **for** $i = 2 : \tau_k - 1$ **do**
            Update $\mathcal{W}_{k,k-i}$ with (5.34).
        **end for**
    **end if**
    Calculate $\mathcal{M}_{k+1}$, $\Sigma_{k+1}$ with (5.26)–(5.27).
**end for**
```

is $\mathcal{Z}_k = 0_{m\times(m-r_k)}$, $[U_k\ \tilde{U}_k] = \hat{U}_k$ *where* $U_k \in \mathbb{R}^{m\times r_k}$ *and* $\hat{Y}_k = \hat{U}_k\hat{S}_k\hat{V}_k^{\mathrm{T}}$ *is the full-size SVD. The design of the optimal attack policy is summarized in Algorithm 5.1.*

Remark 5.7. *From Theorem 5.1, the optimal attack policy can be derived with three steps.* (1) *Obtain the MMSE estimate of* $\hat{e}_k$ *based on* $\mathbb{I}_k$*, denoted by* ϕ_k*;* (2) *de-correlate* ϕ_k *with* $\{\tilde{z}^*_{k-1}, \cdots, \tilde{z}^*_{k-\tau_k+1}\}$ *to obtain* $\bar{\theta}_k$*;* (3) *design* $\tilde{z}^*_k$ *as a linear transformation of* $\bar{\theta}_k$ *added by a compensatory Gaussian noise. The online measurement is utilized only in step 1. This separation principle allows for handling different information scenarios in a unified framework. The "interval uncorrealtion" constraint is tackled in step 2, which provides the freedom for attackers to deceive anomaly detectors with different width. The worst-case estimation performance degradation is determined by both the amount of online information and the width of detection interval.*

The optimal attack has made full utilization of online data $\{\bar{y}_{\bar{k}}, ..., \bar{y}_k\}$ since the state estimation in (5.14) is based on $\mathbb{I}_k$. In practical cases, the attacker needs to design $\tilde{y}^*_k$ firstly, then replace the nominal data with the compromised one. This process can induce transmission delays and may cause the attack being notified by the remote estimator. One solution to tackle the issue is to adopt a one-step ahead predictor instead of (5.14) for state estimation. Then $\tilde{y}^*_k$ is based on the information set $\{\bar{y}_{\bar{k}}, ..., \bar{y}_{k-1}\}$. The design procedure is similar to Theorem 5.1. It is clear that this policy leads to less performance degradation, but has lower level requirement on real-time calculation.

5.2.6 Dynamic Linear Attack Model

In this section, we show that the optimal attack policy in Theorem 5.1 is the output of a dynamic linear system, the model of which is given in the following theorem.

Theorem 5.2. *The optimal compromised measurement is generated by the*

LTV system:

$$\eta_k = A^{\eta}_{k-1}\eta_{k-1} + B^u_k u_k + B^y_k \bar{y}_k, \tag{5.58}$$

$$\tilde{y}^*_k = C^{\eta}_k \eta_k + D^u_k u_k + b_k,\ b_k \sim \mathcal{N}(0_{m\times 1}, \Phi^*_k), \tag{5.59}$$

with the initial condition $\eta_{\bar{k}} = [\alpha^{\mathrm{T}}_{\bar{k}}, \beta^{\mathrm{T}}_{\bar{k}}]^{\mathrm{T}}$ *and coefficient matrices:*

$$A^{\eta}_{k-1} = \begin{bmatrix} A - \bar{K}_k\bar{C}A & 0_n \\ -\bar{K}_k\bar{C}A & A \end{bmatrix},\ B^u_k = \begin{bmatrix} 0_{n\times(\bar{\tau}_k - m)} \\ -AKL_k \end{bmatrix},$$

$$B^y_k = \begin{bmatrix} \bar{K}_k \\ \bar{K}_k \end{bmatrix},\ C^{\eta}_k = \begin{bmatrix} C & H^{\phi}_k K^{\mathrm{T}} - C \end{bmatrix},\ D^u_k = H^z_k,$$

where $H^{\phi}_k \in \mathbb{R}^{m\times m}$, H^z_k *and* L_k *are given by*

$$H^*_k = \begin{bmatrix} H^{\phi}_k & H^z_k \end{bmatrix},\ L_k = \begin{bmatrix} I_m & 0_{m\times(\bar{\tau}_k - 2m)} \end{bmatrix}.$$

The control input is $u_k = [(\tilde{z}^*_{k-1})^{\mathrm{T}}, \cdots, (\tilde{z}^*_{k-\tau_k+1})^{\mathrm{T}}]^{\mathrm{T}}$.

Proof. From (5.14a)–(5.14b), we have

$$\alpha_k = (I_n - \bar{K}_k\bar{C})A\alpha_{k-1} + \bar{K}_k\bar{y}_k. \tag{5.60}$$

Combining (5.60) and (5.12) yields

$$\begin{aligned}\alpha_k - \tilde{x}_{k|k-1} = A(\alpha_{k-1} - \tilde{x}_{k-1|k-2}) - \bar{K}_k\bar{C}A\alpha_{k-1} \\ -AK\tilde{z}^*_{k-1} + \bar{K}_k\bar{y}_k.\end{aligned}$$

Since $\tilde{z}^*_{k-1} = L_k u_k$, it follows that

$$\beta_k = A\beta_{k-1} - \bar{K}_k\bar{C}A\alpha_{k-1} - AKL_k u_k + \bar{K}_k\bar{y}_k. \tag{5.61}$$

Define the state vector $\eta_k = [\alpha^{\mathrm{T}}_k, \beta^{\mathrm{T}}_k]^{\mathrm{T}}$. From (5.20), (5.60)–(5.61) and the facts that $\tilde{y}^*_k = \tilde{z}^*_k + C\tilde{x}_{k|k-1}$, $\tilde{z}^*_k = H^*_k\theta_k + b_k$, and $\tilde{x}_{k|k-1} = \alpha_k - \beta_k$, it is easy to verify that the above conclusion holds. ■

Remark 5.8. *All the involved matrices of the LTV system are independent of measurement data and thus can be determined offline with Algorithm 5.1. In practical cases, to reduce the computational burden, the attacker can store* *the parameters offline and generate the compromised measurement online with* (5.58)–(5.59). *Note that if* $\tau_k = 1$, B_k^u, D_k^u *and* u_k *vanish; then the system is driven only by the online measurement* $\bar{y}_k$. *Theorem 5.2 also indicates the difference of the information-based attack from the innovation-based one, where a static linear model is adopted [22, 39, 68, 69, 97].*

The filter in (5.14) will reduce to a fixed-gain estimator when k tends to infinity; then A_k^η and B_k^y converge to constant matrices. If $\tau \neq \infty$, $\bar{\tau}_k$ and L_k will become constants when $k \geq \bar{k} + \tau - 1$; then B_k^u is a constant. (5.58) reduces to a linear time-invariant system.

5.2.7 Some Special Cases

From the separation principle in Remark 5.7, the attack performance depends on the length of the detection interval and estimation quality for $\tilde{e}_{k|k-1}$. In this section, we discuss some special cases of $\mathbb{I}_k$.

Theorem 5.3. *If* $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k\}$, *an optimal attack policy at the early stage of deception attacks is given by* $\tilde{z}_k^* = -z_k, \forall k \in [\![\bar{k}, \bar{k} + \tau - 1]\!]$.

Proof. We prove by induction. In the case that only the measurement of sensor I is available, we have $\bar{y}_k = y_k, \bar{C} = C, \bar{R} = R$. The Kalman filter in (5.14) is the same as the one in (5.3). Since $\bar{P}_{\bar{k}}^e = \bar{P}$, the filter is in steady state; then $\bar{K}_k = K, P_k^e = (I_n - KC)\bar{P}$, $\forall k \geq \bar{k}$. It follows from (5.14) that

$$\alpha_k = A\alpha_{k-1} + Kz_k. \tag{5.62}$$

Combining (5.62) and (5.12) yields

$$\begin{aligned}\beta_k &= A\beta_{k-1} + Kz_k - AK\tilde{z}_{k-1} \\ &= A^{k-\bar{k}}\beta_{\bar{k}} - \sum_{i=\bar{k}}^{k-1} A^{k-i}K\tilde{z}_i + \sum_{i=\bar{k}+1}^{k} A^{k-i}Kz_i.\end{aligned} \tag{5.63}$$

When the attack starts, we have

$$\beta_{\bar{k}} = \alpha_{\bar{k}} - \tilde{x}_{\bar{k}|\bar{k}-1} = x_{\bar{k}|\bar{k}} - x_{\bar{k}|\bar{k}-1} = Kz_{\bar{k}}. \tag{5.64}$$

Let $\bar{k} < \hat{k} < \bar{k} + \tau$ and assume that

$$\tilde{z}_k^* = -z_k, \forall k \in [\![\bar{k}, \hat{k} - 1]\!]. \tag{5.65}$$

From (5.63)–(5.64) we have

$$\beta_{\hat{k}} = A^{\hat{k}-\bar{k}} K z_{\bar{k}} + \sum_{i=\bar{k}}^{\hat{k}-1} A^{\hat{k}-i} K z_i + \sum_{i=\bar{k}+1}^{\hat{k}-1} A^{\hat{k}-i} K z_i + K z_{\hat{k}}. \tag{5.66}$$

Recall that $\phi_{\hat{k}} = K^{\mathrm{T}} \beta_{\hat{k}}$. To satisfy the "interval uncorrelation" constraint, $\phi_{\hat{k}}$ should be made uncorrelated with all historical compromised innovations in the interval $[\![\bar{k}, \hat{k} - 1]\!]$ to obtain $\bar{\theta}_{\hat{k}}$ (see the proof of Theorem 5.1 and the separation principle in Remark 5.7). According to (5.65)–(5.66), the statement directly leads to $\bar{\theta}_{\hat{k}} = K^{\mathrm{T}} K z_{\hat{k}}$. Then $\tilde{z}_{\hat{k}}$ is designed as

$$\tilde{z}_{\hat{k}} = -\Pi_{\hat{k}} K^{\mathrm{T}} K z_{\hat{k}} + b_{\hat{k}},\ b_{\hat{k}} \sim \mathcal{N}(0_{m\times 1}, \Phi_{\hat{k}}). \tag{5.67}$$

Since $\mathbb{E}[\bar{\theta}_{\hat{k}} \bar{\theta}_{\hat{k}}^{\mathrm{T}}] = P_{\hat{k}}^{\phi} - \mathcal{M}_{\hat{k}} \Sigma_{\hat{k}}^{-1} \mathcal{M}_{\hat{k}}^{\mathrm{T}} = K^{\mathrm{T}} K \Sigma K^{\mathrm{T}} K$, from Theorem 5.1, we have

$$\Pi_{\hat{k}} = -\bar{H}_{\hat{k}} \bar{\Delta}_{\hat{k}} = \Sigma^{\frac{1}{2}} (U_{\hat{k}} S_{\hat{k}}^{-1} U_{\hat{k}}^{\mathrm{T}} + \mathcal{Z}_{\hat{k}} \tilde{U}_{\hat{k}}^{\mathrm{T}}) \Sigma^{\frac{1}{2}}, \tag{5.68}$$

where $\Sigma^{\frac{1}{2}} (K^{\mathrm{T}} K \Sigma K^{\mathrm{T}} K)^{\frac{1}{2}} = U_{\hat{k}} S_{\hat{k}} V_{\hat{k}}^{\mathrm{T}}$. It follows that

$$\Sigma^{\frac{1}{2}} K^{\mathrm{T}} K \Sigma^{\frac{1}{2}} = U_{\hat{k}} S_{\hat{k}} U_{\hat{k}}^{\mathrm{T}}. \tag{5.69}$$

According to (5.68)–(5.69), we have

$$\begin{aligned} -\Pi_{\hat{k}} K^{\mathrm{T}} K &= -\Pi_{\hat{k}} \Sigma^{-\frac{1}{2}} (\Sigma^{\frac{1}{2}} K^{\mathrm{T}} K \Sigma^{\frac{1}{2}}) \Sigma^{-\frac{1}{2}} \\ &= -\Sigma^{\frac{1}{2}} U_{\hat{k}} U_{\hat{k}}^{\mathrm{T}} \Sigma^{-\frac{1}{2}}. \end{aligned} \tag{5.70}$$

If $\mathrm{rank}(C) = m$, $\tilde{U}_{\hat{k}}$ vanishes; then $\Phi_{\hat{k}} = 0_m$, (5.70) implies $-\Pi_{\hat{k}} K^{\mathrm{T}} K = -I_m$; hence (5.67) leads to the unique optimal attack $\tilde{z}_{\hat{k}}^* = -z_{\hat{k}}$. If $\mathrm{rank}(C) < m$, we choose

$$b_{\hat{k}} = -\Sigma^{\frac{1}{2}} \tilde{U}_{\hat{k}} \tilde{U}_{\hat{k}}^{\mathrm{T}} \Sigma^{-\frac{1}{2}} z_{\hat{k}}.$$

With (5.70), we see that $b_{\hat{k}}$ is independent of $-\Pi_{\hat{k}} K^{\mathrm{T}} K z_{\hat{k}}$ and thus is a legitimate choice. It follows from (5.67) that

$$\tilde{z}_{\hat{k}}^* = -\Sigma^{\frac{1}{2}} U_{\hat{k}} U_{\hat{k}}^{\mathrm{T}} \Sigma^{-\frac{1}{2}} z_{\hat{k}} + b_{\hat{k}} = -z_{\hat{k}}. \tag{5.71}$$

The above analysis shows that if (5.65) holds, then $-z_{\hat{k}}$ is an optimal attack at instant $\hat{k}$ (not necessarily unique). Now it is sufficient to prove $-z_{\bar{k}}$ is an optimal attack at instant $\bar{k}$. Since $\phi_{\bar{k}} = K^{\mathrm{T}}\beta_{\bar{k}} = K^{\mathrm{T}}Kz_{\bar{k}}$ and

$$P^{\phi}_{\bar{k}} = K^{\mathrm{T}}P^{\beta}_{\bar{k}}K = K^{\mathrm{T}}(\tilde{P}^{-}_{\bar{k}} - P^{e}_{\bar{k}})K = K^{\mathrm{T}}KCPK.$$

From Theorem 5.1, the optimal attack is given by $\tilde{z}_{\bar{k}} = H^{*}_{\bar{k}}\phi_{\bar{k}} + b_{\bar{k}}$, with the coefficients

$$H^{*}_{\bar{k}} = -\Sigma^{\frac{1}{2}}(U_{\bar{k}}S^{-1}_{\bar{k}}U^{\mathrm{T}}_{\bar{k}} + \mathcal{Z}_{\bar{k}}\tilde{U}^{\mathrm{T}}_{\bar{k}})\Sigma^{\frac{1}{2}},$$

where $\Sigma^{\frac{1}{2}}(K^{\mathrm{T}}KCPK)^{\frac{1}{2}} = U_{\bar{k}}S_{\bar{k}}V^{\mathrm{T}}_{\bar{k}}$. Note that $CP = \Sigma K^{\mathrm{T}}$. Following the same arguments as in (5.67)–(5.71), one can verify that $\tilde{z}^{*}_{\bar{k}} = -z_{\bar{k}}$ is indeed an optimal attack. This completes the proof. ■

Remark 5.9. *If the attacker can only eavesdrop on the original measurement and the compromised innovation is required to be white Gaussian (the width of the anomaly detector is $\tau = \infty$), our results reduce to [22], i.e., flipping the sign of the current nominal innovation is the optimal attack policy. Note that if* $\operatorname{rank}(C) = m$, *$\tilde{z}^{*}_{k} = -z_{k}$ is the unique optimal attack, otherwise there exist multiple optimal solutions.*

Theorem 5.4. *If $\mathbb{I}_k = \emptyset$, the optimal attack policy $\{\tilde{z}^{*}_{k}\}$, $k \in [\![\bar{k}, \bar{k}+\tau-1]\!]$ is a white Gaussian sequence.*

Proof. If the attacker cannot gain any online information, the filter in (5.14) becomes

$$\alpha_k = A\alpha_{k-1},\ P^{e}_{k} = AP^{e}_{k-1}A^{\mathrm{T}} + Q,$$

with the initial condition $\alpha_{\bar{k}} = x_{\bar{k}|\bar{k}-1}$, $P^{e}_{\bar{k}} = \bar{P}$. Following the similar arguments as the proof of Theorem 5.3, one can verify that the conclusion holds. ■

Remark 5.10. *If $\mathbb{I}_k = \emptyset$, the attacker can simply generate a white Gaussian noise in the early stage of deception attacks to design $\tilde{z}_k$. When $k \geq \bar{k} +$*

τ, the information contained in $\{\tilde{z}_{\bar{k}}^*, \cdots, \tilde{z}_{k-\tau}^*\}$ can be utilized to design $\tilde{z}_k^*$, making $\{\tilde{z}_k^*\}$ no longer white. This is the reason that the optimal attack policy in Theorem 5.1 outperforms the intuitive one that $\tilde{z}_k$ is designed as a white Gaussian noise in the whole time horizon.

5.3 Optimal Attacks with Relaxed Stealthiness

In this section, we set $\delta > 0$ and study the optimal deception attacks with relaxed stealthiness. According to (5.13b), the problem is formulated as

$$\begin{aligned} \mathbf{P}_{5.4}: \quad \min_{\tilde{z}_k=\pi_k(\mathbb{I}_k)} \quad & \mathrm{Tr}\{-K\mathbb{E}[\tilde{z}_k\tilde{z}_k^{\mathrm{T}}]K^{\mathrm{T}}\} + 2\,\mathrm{Tr}\{\mathbb{E}[\hat{e}_k\tilde{z}_k^{\mathrm{T}}]\} \\ \text{s.t.} \quad & \mathcal{D}_{\mathrm{KL}}(\tilde{z}_k \| z_k) \le \delta \;\text{ and }\; (5.22). \end{aligned}$$

The optimal attack can also be obtained with the separation principle in Remark 5.7. In this scenario, the linear transformation coefficient in the third step is derived by numerically solving an optimization problem. The main result is given in Theorem 5.5.

Theorem 5.5. *The optimal deception attack policy with relaxed stealthiness is given by*

$$\tilde{y}_k^* = C\tilde{x}_{k|k-1} + H_k^*\theta_k + b_k,\; b_k \sim \mathcal{N}(0_{m\times 1}, \Phi_k^*), \tag{5.72}$$

where H_k^* and Φ_k^* are obtained by solving

$$\begin{aligned} \min_{\tilde{\Sigma}_k, H_k, \Phi_k} \quad & \mathrm{Tr}(-K\tilde{\Sigma}_kK^{\mathrm{T}}) + 2\,\mathrm{Tr}\left\{\begin{bmatrix} P_k^{\phi} & \mathcal{M}_k \end{bmatrix} H_k^{\mathrm{T}}\right\} \\ \text{s.t.} \quad & \mathrm{Tr}(\Sigma^{-1}\tilde{\Sigma}_k) + \ln\frac{|\Sigma|}{|\tilde{\Sigma}_k|} - m - 2\delta \le 0, \\ & H_k \begin{bmatrix} P_k^{\phi} & \mathcal{M}_k \\ \mathcal{M}_k^{\mathrm{T}} & \Sigma_k \end{bmatrix} H_k^{\mathrm{T}} + \Phi_k = \tilde{\Sigma}_k, \\ & H_k \begin{bmatrix} \mathcal{M}_k \\ \Sigma_k \end{bmatrix} = 0_{m\times(\bar{\tau}_k - m)}, \\ & \Phi_k \succeq 0. \end{aligned}$$

P_k^{ϕ} *is given by* (5.18). Σ_k *and* $\mathcal{M}_k$ *are re-defined as*

$$\begin{aligned}\mathcal{M}_k &= \begin{bmatrix}\mathcal{M}_{k,k-1} & \cdots & \mathcal{M}_{k,k-\tau_k+1}\end{bmatrix},\\ \Sigma_k &= \text{blkdiag}\{\tilde{\Sigma}^*_{k-1}, \cdots, \tilde{\Sigma}^*_{k-\tau_k+1}\},\end{aligned}$$

with $\mathcal{M}_{i,j} = K^{\mathrm{T}}\mathcal{W}_{i,j}$. *The recursion of* $\mathcal{W}_{i,j}$ *is given by* (5.34) *and*

$$\begin{aligned}\mathcal{W}_{k,k-1} = {} & A\begin{bmatrix}P^{\beta}_{k-1}K & \mathcal{W}_{k-1,k-2} & \cdots & \mathcal{W}_{k-1,k-\tau_{k-1}}\end{bmatrix}H^{*\mathrm{T}}_{k-1}\\ & -AK\tilde{\Sigma}^*_{k-1}.\end{aligned}$$

Proof. See Appendix A.7. ■

We see that Theorem 5.1 is a special case of the above result when $\delta = 0$. In $\mathbf{P}_{5.4}$, $\tilde{z}_k$ under KL divergence constraint can have an arbitrary probability distribution. Theorem 5.5 shows that the optimal compromised innovation is zero-mean Gaussian. The filter under deception attacks still provides unbiased state estimation. The attack performance is now evaluated by

$$\begin{aligned}\tilde{P}_{k|k} = {} & \tilde{P}_{k|k-1} + K\Sigma^*_k K^{\mathrm{T}} - KH^*_k\begin{bmatrix}P^{\phi}_k & \mathcal{M}_k\end{bmatrix}^{\mathrm{T}}K^{+}\\ & -(K^{\mathrm{T}})^{+}\begin{bmatrix}P^{\phi}_k & \mathcal{M}_k\end{bmatrix}(H^*_k)^{\mathrm{T}}K^{\mathrm{T}}.\end{aligned} \tag{5.73}$$

The design of optimal attacks with relaxed stealthiness follows a similar procedure as Algorithm 5.1. Accordingly, $\tilde{y}^*_k$ can also be generated by an LTV system.

5.4 Examples

Numerical examples are provided in this section to verify the theoretical results. A stable LTI system is given with the following parameters:

$$A = \begin{bmatrix}0.717 & -0.043 & -0.082\\ -0.043 & 0.666 & -0.025\\ -0.082 & -0.025 & 0.718\end{bmatrix}, Q = \text{diag}\left\{\begin{bmatrix}0.612\\ 0.435\\ 0.754\end{bmatrix}\right\}$$

$$C = \begin{bmatrix}1.326 & 0.756 & 2.352\\ -1.319 & 0.921 & 0.395\\ 0.896 & 1.564 & -1.887\end{bmatrix}, R = \text{diag}\left\{\begin{bmatrix}1.054\\ 2.026\\ 1.648\end{bmatrix}\right\}$$

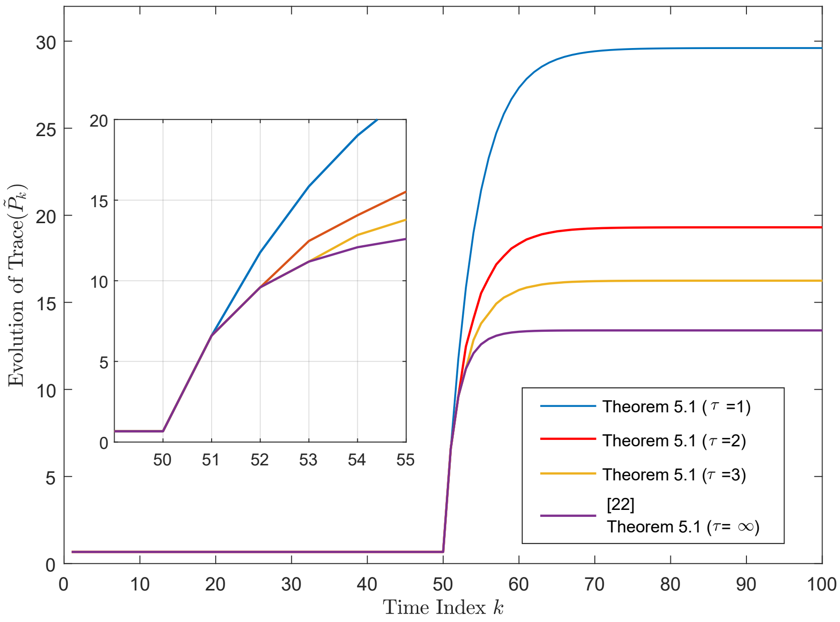

Figure 5.3: Performance evolution of optimal stealthy deception attacks ($\delta = 0$) with different χ^2 detectors.

Assume $\bar{k} = 51$, the attacker can only eavesdrop on the original measurements ($\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k\}$). The performance of optimal deception attacks that can completely bypass different χ^2 detectors is illustrated in Fig. 5.3. It shows that the worst-case attack impact can be mitigated by increasing the width of the detection window. $\tau = 1$ and $\tau = \infty$ correspond to the performance limits of two extreme cases. The simulation also justifies the necessity to implement multi-step χ^2 detectors in practical systems. Compared with a single-step detector, the multi-step one is not only more robust to disturbances but also reduces the attack performance by imposing a more strict constraint for attackers. The enlarged figure illustrates the effectiveness of Theorem 5.3, i.e., $\tilde{z}_k^* = -z_k$ is the (unique) optimal attack when $k < \bar{k} + \tau$.

The performance of optimal innovation-based and information-based attacks is illustrated in Fig. 5.4. Both the attacks in Theorem 5.1 ($\tau = 1$) and [68] using 4 historical nominal innovations can only deceive a single-step χ^2 detector. Note that when $k \leq 55$, these two attacks have the same perfor-

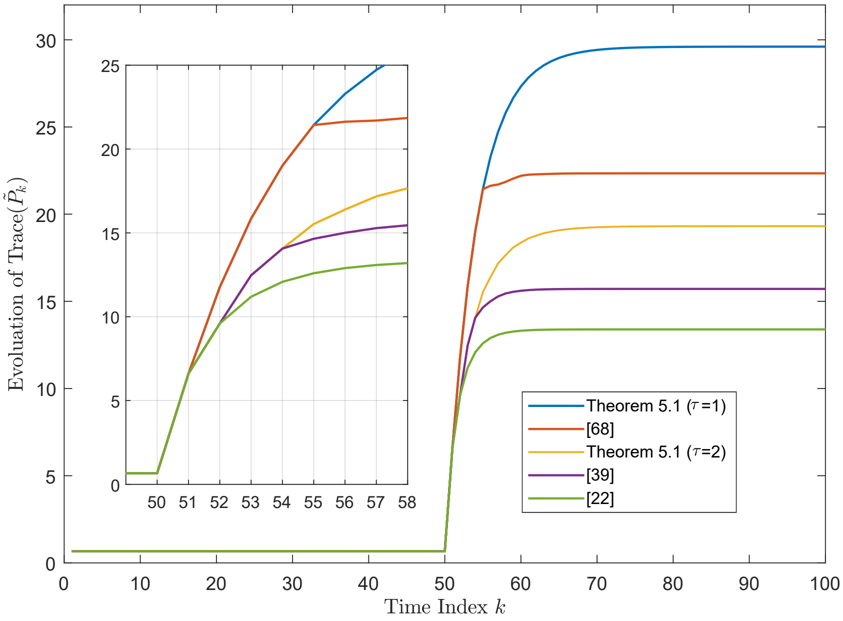

Figure 5.4: Performance comparison of innovation-based linear attacks and information-based attacks.

mance because all information is fully utilized. When $k > 55$, the latter one is no longer optimal. Both the attacks in Theorem 5.1 ($\tau = 2$) and [39] can deceive a 2-step detector. The figure shows that the information-based attack can cause more estimation quality degradation compared with the innovation-based ones.

We conduct Monte Carlo simulations for 5,000 times to show the stealthiness property. Assume $\bar{k} = 301$, the detection window is of size 2 and $J_{th} = 10$, leading to a theoretical FAR $\alpha = 12.47\%$. Fig. 5.5 indicates that the stealthy deception attack from Theorem 5.1 with $\tau = 2$ can completely deceive a 2-step detector. If the attacker mistakenly uses $\tau = 1$ to design the optimal attack, Fig. 5.6 shows that the attack is no longer strictly stealthy as a higher alarm rate is induced.

We then use the stable LTI system in [24] to illustrate the impact of side information ($\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k, \hat{y}_k\}$). Let $\bar{k} = 31$, $S = 0_2$ and

$$A = \begin{bmatrix} 0.8 & 0.6 & 0 \\ 0 & 0.5 & 0.3 \\ 0 & 0 & 0.7 \end{bmatrix}, \; C = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \; \hat{C} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

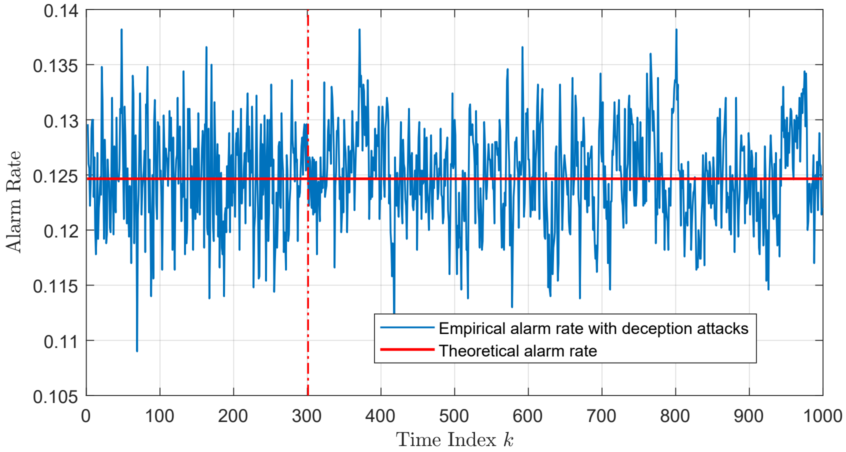

Figure 5.5: The optimal attack in Theorem 5.1 ($\tau = 2$) can completely bypass a 2-step anomaly detector.

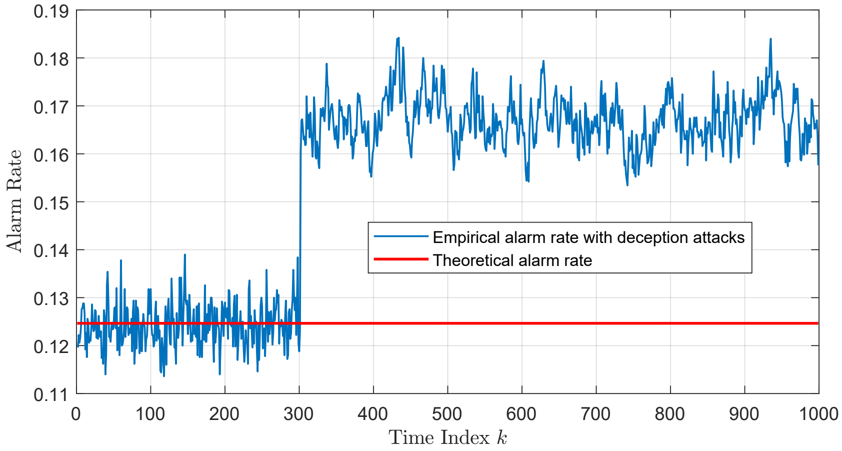

Figure 5.6: The optimal attack in Theorem 5.1 ($\tau = 1$) cannot deceive a 2-step anomaly detector.

Fig. 5.7 shows that additional information can *always* improve the attack performance. Since the attacks in [24] and [97] were based on the MMSE estimate of x_k but not $\tilde{e}_{k|k-1}$, these policies have not utilized the online information properly and thus did not achieve the maximum performance loss. We also see that the attack performance depends on both the amount of available information and the width of detectors.

Let $\delta = 0.1$. Fig. 5.8 illustrates the performance of attacks with relaxed stealthiness. It is clear that the attacker can achieve greater performance degradation by sacrificing the stealthiness property. The information-based

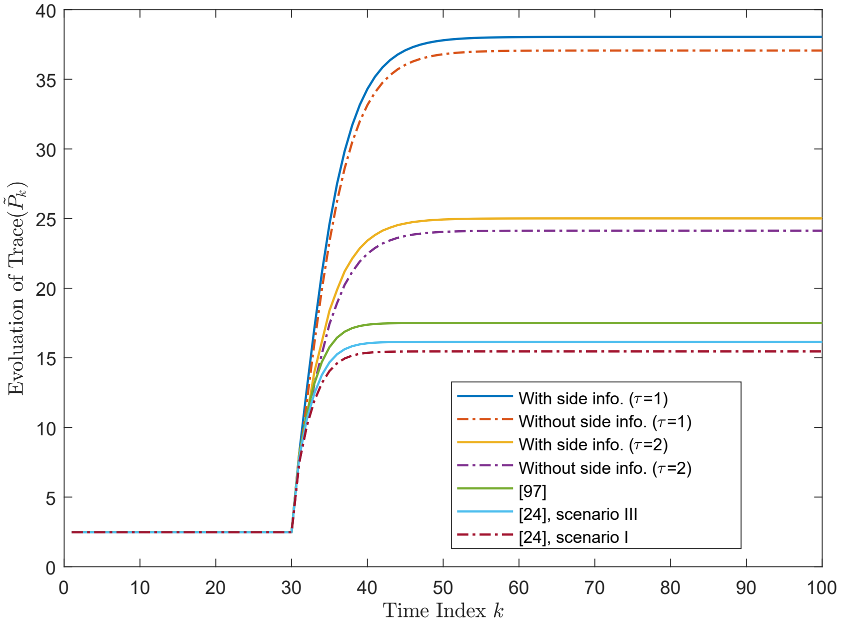

Figure 5.7: Attack performance with/without side information.

policy also outperforms the innovation-based one with the same stealthiness level.

5.5 Conclusions

In this chapter, the problem of deception attacks on remote state estimators equipped with interval anomaly detectors has been solved completely. The major challenge is how to properly handle the stealthiness constraint such that the attacker's behavior is not overly restricted. A separation principle consisting of three steps is proposed to design the information-based deception policy. The attack performance depends on both the amount of online information and the width of the χ^2 detector. Contrary to existing studies that assume innovation-based (static) linear attacks, the information-based one is shown to be generated by a dynamic LTV system. The result reduces to optimal innovation-based attacks in some special cases.

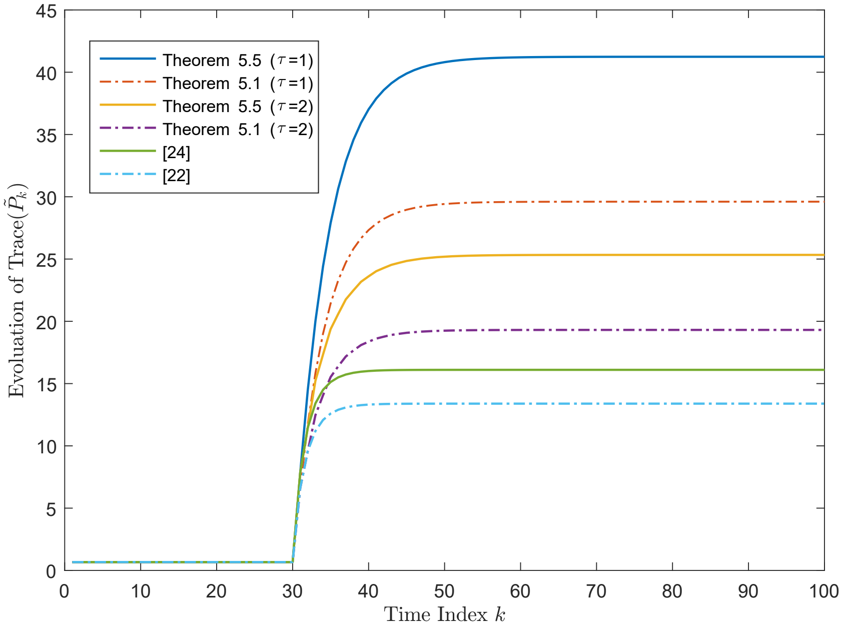

Figure 5.8: Attack performance with relaxed/strict stealthiness.

Chapter 6

Optimal Information-Based Deception Attacks with Maximum Interval Performance *

This chapter studies the problem of optimal deception attacks against remote state estimation, where the measurement data is transmitted through an unreliable wireless channel. A malicious agent is capable of intercepting and modifying raw data, with the goal to cause the maximum estimation quality degradation and deceive an interval χ^2 detector. Contrary to existing studies that focused on greedy attack performance, we consider a more general scenario that the attacker aims to maximize the summation of estimation errors in a fixed interval. It is shown that the information-based optimal attack is a linear combination of MMSE estimates of all historical prediction errors. The combination coefficients can be obtained by solving a convex optimization problem. The effectiveness of the method is verified with numerical examples and comparative study with existing work.

This chapter is organized as follows. Section 6.1 gives the detailed problem description. The optimal attack structure and coefficients are derived in

*A preliminary version of this chapter has been submitted for publication as: Jing Zhou, Jun Shang, and Tongwen Chen, Deception attacks on Kalman filtering with interval estimation performance loss, *5th IFAC Workshop on Linear Parameter Varying Systems*, Montreal, Sept. 2022. (Accepted)

Section 6.2. Section 6.3 discusses some special cases when multiple-step χ^2 detectors and whiteness detectors are deployed to reveal anomalies. Numerical examples are given in Section 6.3, followed by conclusions in Section 6.4.

6.1 Problem Formulation

The system architecture is illustrated in Fig. 4.1. In this Chapter, we consider the same process model, remote state estimator, anomaly detector, and cyber-attacks as in Chapter 4. The attack impact at each step can be evaluated by $\mathrm{Tr}(\tilde{P}_{k|k})$. The problem of concern is to design an optimal stealthy attack sequence $\mathbb{I}_k$, such that the compromised innovation $\{\tilde{z}_k\}$ satisfies $\tilde{z}_k \sim \mathcal{N}(0_m, \Sigma)$ and the following performance index in $[\![\bar{k}, k]\!]$ is maximized:

$$J_{[\bar{k},k]} = \sum_{i=\bar{k}}^{k} \mathrm{Tr}\, \tilde{P}_{i|i}. \tag{6.1}$$

Remark 6.1. *Compared with maximizing* $\mathrm{Tr}(\tilde{P}_{i|i})$ *directly step by step, the optimization of* $J_{[\bar{k},k]}$ *is a more challenging task. The main difficulty is that the attack effect at each step will propagate through the evolution of the filter dynamics. Instead of analyzing the step-wise correlation of* $\tilde{z}_k$ *and prediction error, the attack effects of* $\tilde{z}_k$ *in the entire time horizon should be considered as a whole for interval performance optimization.*

6.2 Preliminaries

With presence of deception attacks, the compromised Kalman filter becomes

$$\tilde{x}_{k|k-1} = A\tilde{x}_{k-1|k-1}, \tag{6.2}$$

$$\tilde{x}_{k|k} = \tilde{x}_{k|k-1} + K\tilde{z}_k. \tag{6.3}$$

The dynamic of the prediction error is given by

$$\tilde{e}_{k|k-1} = A\tilde{e}_{k-1|k-2} - AK\tilde{z}_{k-1} + w_{k-1}. \tag{6.4}$$

From (6.2)–(6.3), we see that $\tilde{x}_{k|k-1}$ is determined by $\tilde{x}_{\bar{k}|\bar{k}-1}$ and $\tilde{z}_i, i \in [\![\bar{k}, k-1]\!]$; thus it is a known variable at instant k. According to (4.16), the "virtual measurement" for $\tilde{e}_{k|k-1}$ can be defined as

$$r_k \triangleq \bar{y}_k - \bar{C}\tilde{x}_{k|k-1} = \bar{C}\tilde{e}_{k|k-1} + \bar{v}_k. \tag{6.5}$$

The attacker adopts the following Kalman filter to obtain an MMSE estimate for $\tilde{e}_{k|k-1}$:

$$\bar{\theta}_k = A\theta_{k-1} - AK\tilde{z}_{k-1}, \tag{6.6a}$$

$$\theta_k = \bar{\theta}_k + \bar{K}_k(r_k - \bar{C}\bar{\theta}_k), \tag{6.6b}$$

$$\bar{K}_k = \bar{P}^e_k\bar{C}^{\mathrm{T}}\left(\bar{C}\bar{P}^e_k\bar{C}^{\mathrm{T}} + \bar{R}\right)^{-1}, \tag{6.6c}$$

$$\bar{P}^e_k = AP^e_{k-1}A^{\mathrm{T}} + Q, \tag{6.6d}$$

$$P^e_k = \left(I_n - \bar{K}_k\bar{C}\right)\bar{P}^e_k, \tag{6.6e}$$

with initial condition $\bar{\theta}_{\bar{k}} = 0_n$, $\bar{P}_{\bar{k}} = \bar{P}$. According to (6.5) and (6.6a)–(6.6b), we have

$$\begin{aligned}\theta_k = {} & (I_n - \bar{K}_k\bar{C})A\theta_{k-1} - AK\tilde{z}_{k-1} \\ & + \bar{K}_k\bar{C}A\tilde{e}_{k-1|k-2} + \bar{K}_k\bar{C}w_{k-1} + \bar{K}_k\bar{v}_k.\end{aligned} \tag{6.7}$$

Using the orthogonality property of MMSE estimation, we know that the covariance of θ_k is given by

$$P^\theta_k = \mathbb{E}[\theta_k\theta_k^{\mathrm{T}}] = \tilde{P}_{k|k-1} - P^e_k. \tag{6.8}$$

The following lemma establishes the connection between an information-based constrained optimization problem and MMSE estimation.

Lemma 6.1. *Let $S \in \mathbb{R}^{n\times m}$, $\Pi \in \mathbb{S}^n_+$, $\Sigma \in \mathbb{S}^m_{++}$, and $\mathcal{X}$ be a random Gaussian vector $\mathcal{X} \sim \mathcal{N}(0_n, \Pi)$. The observed information set for $\mathcal{X}$ is denoted as $\mathbb{I}_x$; then a vector that satisfies $\mathcal{Y} \sim \mathcal{N}(0_m, \Sigma)$ and also minimizes $\mathrm{Tr}\{S\mathbb{E}[\mathcal{Y}\mathcal{X}^{\mathrm{T}}]\}$ is given by*

$$\mathcal{Y}^* = L_x\hat{\mathcal{X}} + b_x,\ b_x \sim \mathcal{N}(0_m, \Phi_x),$$

where $\hat{\mathcal{X}} = \mathbb{E}[\mathcal{X}|\mathbb{I}_x]$ *is the MMSE estimate of* $\mathcal{X}$*,* b_x *is a Gaussian noise, and the coefficients are given as*

$$L_x = -\mathcal{R}^{+}S^{\mathrm{T}},$$
$$\Phi_x = \Sigma - L_x\mathcal{P}_s L_x^{\mathrm{T}},$$

with $\mathcal{R} = \Sigma^{-\frac{1}{2}}(\Sigma^{\frac{1}{2}}\mathcal{P}_s\Sigma^{\frac{1}{2}})^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}$*,* $\mathcal{P}_s = S^{\mathrm{T}}(\Pi - \mathcal{P}_e)S$*, and* $\mathcal{P}_e$ *the estimation error covariance:*

$$\mathcal{P}_e = \mathbb{E}[(\mathcal{X} - \hat{\mathcal{X}})(\mathcal{X} - \hat{\mathcal{X}})^{\mathrm{T}}|\mathbb{I}_x].$$

Proof. See Appendix A.8. ■

An example is provided to briefly illustrate the intuition in Lemma 6.1. Assume $x \sim \mathcal{N}(0,1)$. z_1 and z_2 are measurements of x corrupted by Gaussian noises:

$$z_1 = 2x + w_1,\ w_1 \sim \mathcal{N}(0,1),$$
$$z_2 = 3x + w_2,\ w_2 \sim \mathcal{N}(0,3),$$

then the random variable that satisfies $y \sim \mathcal{N}(0,4)$ and also minimizes $\mathbb{E}[xy]$ is given by

$$y^{*} = L_x\mathbb{E}[x|z_1,z_2] = -0.5345z_1 - 0.2673z_2.$$

In this example we have $\Phi_x = 0$, indicating that the Gaussian noise is not needed to satisfy the constraint. Note that $\mathcal{Y}^{*}$ is not unique when $\mathcal{P}_s$ is singular. Multiple optimal solutions can be obtained by selecting different free parameters $\hat{\mathcal{Z}}$ and $\bar{\mathcal{Z}}$. All of them achieves the same minimum objective value. The details on the uniqueness of $\mathcal{Y}^{*}$ are discussed in the proof of Lemma 6.1. Specifically, if $\mathrm{rank}(\mathcal{P}_s) = m$, we have $L_x = \mathcal{R}^{-1}S^{\mathrm{T}}$, $\Phi_x = 0_m$. The compensatory noise term vanishes.

The optimal deception attack that can completely deceive single-step χ^2 detectors and achieve maximum interval performance is given in the next section.

6.3 Main Results

We first study the attack performance. Let $s = k - \bar{k} + 1$ denote the length of attack horizon. Define $\hat{P} = (I_n - KC)\bar{P}$, $\hat{\Sigma} = K\Sigma K^{\mathrm{T}} + Q$ and the following matrices

$$Y_i = \sum_{j=0}^{k-i} (A^{\mathrm{T}})^j A^j, \ i \in [\![\bar{k}, k]\!].$$

Lemma 6.2. *The interval attack performance in* (6.1) *is evaluated by* $J_{[\bar{k},k]} = -2\hat{J}_{[\bar{k},k]} + \mathrm{Tr}(\mathcal{C})$*, where*

$$\hat{J}_{[\bar{k},k]} = \sum_{i=\bar{k}}^{k} \mathrm{Tr}\{Y_i K \mathbb{E}[\tilde{z}_i \tilde{e}_{i|i-1}^{\mathrm{T}}]\} \tag{6.9}$$

and $\mathcal{C}$ *is a constant matrix given by*

$$\mathcal{C} = \sum_{i=1}^{s} \sum_{j=0}^{i-1} A^j \hat{\Sigma} (A^{\mathrm{T}})^j + \sum_{i=1}^{s} A^i \hat{P} (A^{\mathrm{T}})^i.$$

Proof. See Appendix A.9. ■

According to Lemma 6.2, we see that maximizing the interval performance is equivalent to minimizing $\hat{J}_{[\bar{k},k]}$. The problem of concern is formulated as

$$\begin{aligned} \mathbf{P}_{6.1}: \quad & \min_{\tilde{z}_k = \pi(\mathbb{I}_k)} \quad \hat{J}_{[\bar{k},k]} \\ & \text{s.t.} \ \ \tilde{z}_i \sim \mathcal{N}(0_m, \Sigma), \ \forall i \in [\![\bar{k}, k]\!]. \end{aligned}$$

6.3.1 Optimal Attack Structure

The information-based attack in $\mathbf{P}_{6.1}$ can have an arbitrary form. In this section we show that the optimal policy minimizing $\hat{J}_{[\bar{k},k]}$ is a linear transformation of the MMSE estimate for $\tilde{e}_{k|k-1}$. The main result is summarized in the following theorem.

Theorem 6.1. *The optimal compromised innovation at instant* $i \in [\![\bar{k}, k]\!]$ *is given by*

$$\tilde{z}_i^* = L_i \theta_i + b_i, \ b_i \sim \mathcal{N}(0_m, \Phi_i), \tag{6.10}$$

the coefficient matrices satisfy $L_i = -\mathcal{R}_i^+ \mathcal{M}_i^{\mathrm{T}}, \Phi_i = \Sigma - L_i \mathcal{P}_i L_i^{\mathrm{T}}$ *where* $\mathcal{R}_i = \Sigma^{-\frac{1}{2}}(\Sigma^{\frac{1}{2}}\mathcal{P}_i\Sigma^{\frac{1}{2}})^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}$, $\mathcal{P}_i = \mathcal{M}_i^{\mathrm{T}} P_i^{\theta} \mathcal{M}_i$ *and*

$$\mathcal{M}_{i-1} = Y_{i-1}K - A^{\mathrm{T}}\mathcal{M}_i L_i AK - A^{\mathrm{T}} L_i^{\mathrm{T}} \mathcal{M}_i^{\mathrm{T}} AK,$$

with initial condition $\mathcal{M}_k = K$.

Proof. Let $t \in \mathbb{N}$ and satisfy $\bar{k} < t \leq k$. Assume $\forall i \in [\![t, k]\!]$, (6.10) holds and the corresponding optimal cost-to-go function has the form:

$$\hat{J}^*_{[i,k]} = \mathrm{Tr}\{\mathcal{M}_i \mathbb{E}[\tilde{z}_i^* \tilde{e}_{i|i-1}^{\mathrm{T}}]\} + \mathcal{C}_i, \tag{6.11}$$

where $\mathcal{M}_i \in \mathbb{R}^{n\times m}$ and $\mathcal{C}_i \in \mathbb{R}$ are constant parameters. We first consider $\hat{J}_{[k,k]}$, i.e., the attack performance in the last step. Note that $Y_k = I_n$; we have

$$\hat{J}_{[k,k]} = \mathrm{Tr}\{K\mathbb{E}[\tilde{z}_k \tilde{e}_{k|k-1}^{\mathrm{T}}]\}.$$

From Lemma 6.1, the optimal attack at instant k is

$$\tilde{z}_k^* = L_k\theta_k + b_k,\ b_k \sim \mathcal{N}(0_m, \Phi_k),$$

where $L_k = -\mathcal{R}_k^+ K^{\mathrm{T}}, \Phi_k = \Sigma - L_k\mathcal{P}_k L_k^{\mathrm{T}}$; thus (6.10) and (6.11) hold for $i = k$ with $\mathcal{M}_i = K$ and $\mathcal{C}_i = 0$. We now consider $\hat{J}_{[t-1,k]}$:

$$\begin{aligned}\hat{J}^*_{[t-1,k]} =& \min_{\tilde{z}_{t-1},\cdots,\tilde{z}_k} \{\hat{J}_{[t,k]} + \mathrm{Tr}[Y_{t-1}K\mathbb{E}(\tilde{z}_{t-1}\tilde{e}_{t-1|t-2}^{\mathrm{T}})]\}\\ =&\hat{J}^*_{[t,k]} + \min_{\tilde{z}_{t-1}} \mathrm{Tr}[Y_{t-1}K\mathbb{E}(\tilde{z}_{t-1}\tilde{e}_{t-1|t-2}^{\mathrm{T}})]\\ =&\min_{\tilde{z}_{t-1}} \mathrm{Tr}\{\mathcal{M}_t\mathbb{E}[\tilde{z}_t^*\tilde{e}_{t|t-1}^{\mathrm{T}}] + Y_{t-1}K\mathbb{E}[\tilde{z}_{t-1}\tilde{e}_{t-1|t-2}^{\mathrm{T}}]\} + \mathcal{C}_t,\end{aligned} \tag{6.12}$$

where the last equality is from the assumption in (6.11). Since $\tilde{z}_t^* = L_t\theta_t + b_t$ and b_t is independent of other variables, it follows from (6.4) and (6.7) that

$$\begin{aligned}\mathbb{E}[\tilde{z}_t^*\tilde{e}_{t|t-1}^{\mathrm{T}}] =&\mathbb{E}[(L_t\theta_t + b_t)(A\tilde{e}_{t-1|t-2} - AK\tilde{z}_{t-1} + w_{t-1})^{\mathrm{T}}]\\ =&L_t\mathbb{E}\{[(I_n - \bar{K}_t\bar{C})A\theta_{t-1} - AK\tilde{z}_{t-1} + \bar{K}_t\bar{C}A\tilde{e}_{t-1|t-2}\\ &+ \bar{K}_t\bar{C}w_{t-1} + \bar{K}_t\bar{v}_t](A\tilde{e}_{t-1|t-2} - AK\tilde{z}_{t-1} + w_{t-1})^{\mathrm{T}}\}.\end{aligned}$$

Note that $\tilde{z}_{t-1}$ is designed based on $\mathbb{I}_{k-1}$ and thus is independent of $w_{t-1}, \bar{v}_t$; the noises are mutually independent. Using the following equalities

$$\begin{aligned}
&\mathbb{E}[\theta_{t-1}\tilde{e}^{\mathrm{T}}_{t-1|t-2}] = P^{\theta}_{t-1},\ \mathbb{E}[\tilde{e}_{t-1|t-2}\tilde{e}^{\mathrm{T}}_{t-1|t-2}] = \bar{P}^{e}_{t-1},\\
&\mathbb{E}[(\tilde{e}_{t-1|t-2} - \theta_{t-1})\tilde{e}^{\mathrm{T}}_{t-1|t-2}] = \bar{P}^{e}_{t-1} - P^{\theta}_{t-1},\\
&\mathbb{E}[\tilde{z}_{t-1}\tilde{z}^{\mathrm{T}}_{t-1}] = \Sigma,\ \mathbb{E}[w_{t-1}w^{\mathrm{T}}_{t-1}] = Q,\\
&\mathbb{E}[\theta_{t-1}\tilde{z}^{\mathrm{T}}_{t-1}] = \mathbb{E}[\tilde{e}_{t-1|t-2}\tilde{z}^{\mathrm{T}}_{t-1}],
\end{aligned}$$

we can verify that

$$\begin{aligned}
\mathbb{E}[\tilde{z}^{*}_{t}\tilde{e}^{\mathrm{T}}_{t|t-1}] =& - L_tAK\tilde{E}[\tilde{z}_{t-1}\tilde{e}^{\mathrm{T}}_{t-1|t-2}]A^{\mathrm{T}}\\
&- L_tA\mathbb{E}[\tilde{e}_{t-1|t-2}\tilde{z}^{\mathrm{T}}_{t-1}]K^{\mathrm{T}}A^{\mathrm{T}} + \hat{\mathcal{C}}_{t-1},
\end{aligned}$$

where $\hat{\mathcal{C}}_{t-1}$ is independent of $\tilde{z}_{t-1}$ and given by

$$\begin{aligned}
\hat{\mathcal{C}}_{t-1} =& L_tAP^{\theta}_{t-1}A^{\mathrm{T}} + L_tAK\Sigma K^{\mathrm{T}}A^{\mathrm{T}} + L_t\bar{K}_t\bar{C}Q\\
&+ L_t\bar{K}_t\bar{C}A(\bar{P}^{e}_{t-1} - P^{\theta}_{t-1})A^{\mathrm{T}}.
\end{aligned}$$

It follows from (6.12) that

$$\hat{J}^{*}_{[t-1,k]} = \min_{\tilde{z}_{t-1}} \mathrm{Tr}\{\mathcal{M}_{t-1}\mathbb{E}[\tilde{z}_{t-1}\tilde{e}^{\mathrm{T}}_{t-1|t-2}]\} + \mathcal{C}_{t-1},$$

where

$$\begin{aligned}
&\mathcal{M}_{t-1} = Y_{t-1}K - A^{\mathrm{T}}\mathcal{M}_tL_tAK - A^{\mathrm{T}}L^{\mathrm{T}}_t\mathcal{M}^{\mathrm{T}}_tAK,\\
&\mathcal{C}_{t-1} = \mathrm{Tr}(\mathcal{M}_t\hat{\mathcal{C}}_{t-1}) + \mathcal{C}_t.
\end{aligned}$$

According to Lemma 6.1, the optimal attack at instant $t-1$ is given by

$$\tilde{z}^{*}_{t-1} = L_{t-1}\theta_{t-1} + b_{t-1},\ b_{t-1} \sim \mathcal{N}(0_m, \Phi_{t-1}),$$

where $L_{t-1} = -\mathcal{R}^{+}_{t-1}\mathcal{M}^{\mathrm{T}}_{t-1}, \Phi_{t-1} = \Sigma - L_{t-1}\mathcal{P}_{t-1}L^{\mathrm{T}}_{t-1}$. Repeating the above argument from $i = k$ to $\bar{k}$ completes the proof. ■

Remark 6.2. *In this section, we utilize dynamic programming to show that the optimal attack at each instant is a linear function of θ_k added by a Gaussian noise. Though the optimality has been established, it is a challenging* *task to find these transformation matrices and covariances of b_i. The main difficulty lies in that P_i^θ is given by* (6.8)*, which is determined by all previous attack coefficients $L_t, t \in [\![\bar{k}, i-1]\!]$, but the recursion of $\mathcal{M}_i$ in Theorem 6.1 is backward from k to $\bar{k}$. This leads to a dilemma: in order to find L_i, one has to know $L_t, t \in [\![\bar{k}, i-1]\!]$, but the later ones can only be obtained after L_i is available. Since L_i and Φ_i can not be calculated in a recursive manner, one has to find the matrices that satisfy these coupled equations simultaneously. These attack coefficients are not unique if* $\min\{\mathrm{rank}(\mathcal{P}_i), \mathrm{rank}(\mathcal{M}_i)\} < m$*. Nevertheless, the most important finding in Theorem 6.1 is that the optimal attack has a particular structure. In next sections we will show that the attack coefficients can be obtained by solving a convex optimization problem.*

6.3.2 Augmented Vectors of $\tilde{e}_{k|k-1}$ and θ_k

We firstly define $e = \tilde{e}_{\bar{k}|\bar{k}-1}$ and the following vectors:

$$\begin{aligned}
\mathcal{E} &= \begin{bmatrix} \tilde{e}_{\bar{k}|\bar{k}-1}^{\mathrm{T}} & \tilde{e}_{\bar{k}+1|\bar{k}}^{\mathrm{T}} & \cdots & \tilde{e}_{k+1|k}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{n(s+1)}, \\
\Theta &= \begin{bmatrix} \theta_{\bar{k}}^{\mathrm{T}} & \theta_{\bar{k}+1}^{\mathrm{T}} & \cdots & \theta_{k+1}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{n(s+1)}, \\
\mathcal{Z} &= \begin{bmatrix} \tilde{z}_{\bar{k}}^{\mathrm{T}} & \tilde{z}_{\bar{k}+1}^{\mathrm{T}} & \cdots & \tilde{z}_{k}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{ms}, \\
\mathcal{W} &= \begin{bmatrix} w_{\bar{k}}^{\mathrm{T}} & w_{\bar{k}+1}^{\mathrm{T}} & \cdots & w_{k}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{ns}, \\
\mathcal{V} &= \begin{bmatrix} \bar{v}_{\bar{k}}^{\mathrm{T}} & \bar{v}_{\bar{k}+1}^{\mathrm{T}} & \cdots & \bar{v}_{k+1}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{(m+\bar{m})(s+1)}.
\end{aligned}$$

In the considered attack horizon, all uncertainties in the system come from e, $\mathcal{W}$ and $\mathcal{V}$. Their covariances are respectively given by

$$\begin{aligned}
\mathcal{P}_e &= \mathbb{E}[ee^{\mathrm{T}}] = \bar{P}, \\
\mathcal{P}_w &= \mathbb{E}[\mathcal{W}\mathcal{W}^{\mathrm{T}}] = \mathrm{blkdiag}(Q, s), \\
\mathcal{P}_v &= \mathbb{E}[\mathcal{V}\mathcal{V}^{\mathrm{T}}] = \mathrm{blkdiag}(\bar{R}, s+1),
\end{aligned}$$

From (6.4) we have

$$\tilde{e}_{k|k-1} = A^{k-\bar{k}}\tilde{e}_{\bar{k}|\bar{k}-1} - \sum_{i=\bar{k}}^{k-1} A^{k-i}K\tilde{z}_i + \sum_{i=\bar{k}}^{k-1} A^{k-1-i}w_i.$$

It then can be verified that the augmented vector $\mathcal{E}$ can be written as

$$\mathcal{E} = H\mathcal{Z} + G\mathcal{W} + Ee, \tag{6.13}$$

where the constant matrices are given by

$$H_{[n,m]}(i,j) = \begin{cases} 0_{n\times m}, \ j \geq i \\ -AK, \ j = i-1 \\ AH_{[n,m]}(i-1,j), \ j < i-1 \end{cases}$$

$$G_{[n,n]}(i,j) = \begin{cases} 0_{n}, \ j \geq i \\ I_n, \ j = i-1 \\ AG_{[n,n]}(i-1,j), \ j < i-1 \end{cases}$$

$$E_{[n,n]}(i,1) = \begin{cases} I_n, \ i = 1 \\ AE_{[n,n]}(i-1,1), \ i > 1 \end{cases}$$

Now define $\bar{A}_k = (I_n - \bar{K}_k\bar{C})A$ and substitute the solution of (6.4) into (6.7). When $k \geq \bar{k}+2$, we have

$$\begin{aligned} \theta_k =& \bar{A}_k\theta_{k-1} + \bar{K}_k\bar{C}A^{k-\bar{k}}\tilde{e}_{\bar{k}|\bar{k}-1} - AK\tilde{z}_{k-1} + \bar{K}_k\bar{v}_k \\ & - \bar{K}_k\bar{C}\sum_{i=\bar{k}}^{k-2} A^{k-i}K\tilde{z}_i + \bar{K}_k\bar{C}\sum_{i=\bar{k}}^{k-1} A^{k-1-i}w_i. \end{aligned} \tag{6.14}$$

From (6.6b), when $k = \bar{k}$, the initial condition of θ_k is

$$\theta_{\bar{k}} = \bar{\theta}_{\bar{k}} + \bar{K}_{\bar{k}}(r_{\bar{k}} - \bar{C}\bar{\theta}_{\bar{k}}) = \bar{K}_{\bar{k}}\bar{C}\tilde{e}_{\bar{k}|\bar{k}-1} + \bar{K}_{\bar{k}}\bar{v}_{\bar{k}},$$

when $k = \bar{k}+1$, we have

$$\begin{aligned} \theta_{\bar{k}+1} =& \bar{A}_{\bar{k}+1}\theta_{\bar{k}} - AK\tilde{z}_{\bar{k}} + \bar{K}_{\bar{k}+1}\bar{C}A\tilde{e}_{\bar{k}|\bar{k}-1} \\ & + \bar{K}_{\bar{k}+1}\bar{C}w_{\bar{k}} + \bar{K}_{\bar{k}+1}\bar{v}_{\bar{k}+1}. \end{aligned}$$

It is clear that the LTV system in (6.14) admits a solution that is a linear combination of $\tilde{z}_k$, w_k, $\bar{v}_k$, and $\tilde{e}_{\bar{k}|\bar{k}-1}$. Therefore, Θ can be written in a compact form:

$$\Theta = \bar{H}\mathcal{Z} + \bar{G}\mathcal{W} + \bar{M}\mathcal{V} + \bar{E}e, \tag{6.15}$$

where the constant matrices are constructed as follows:

(*i*) $\bar{H}_{[n,m]}(i,j)$ denotes the transformation matrix from $\tilde{z}_{\bar{k}+j-1}$ to $\theta_{\bar{k}+i-1}$. Since θ_i is independent of $\tilde{z}_j, \forall j \in [\![i,k]\!]$, we have $\bar{H}_{[n,m]}(i,j) = 0_{n\times m}, \forall j \geq i$. When $i = j+1$, the linear subsystem where $\tilde{z}_{\bar{k}+i-2}$ serves as the input in (4.31) is extracted as:

$$\theta^z_{\bar{k}+i-1} = \bar{A}_{\bar{k}+i-1}\theta^z_{\bar{k}+i-2} - AK\tilde{z}_{\bar{k}+i-2},$$

thus $\bar{H}_{[n,m]}(i,i-1) = -AK, \forall i \in [\![2,s+1]\!]$. Similarly, when $i \geq j+2$, from (6.14) we have

$$\theta^z_{\bar{k}+i-1} = \bar{A}_{\bar{k}+i-1}\theta^z_{\bar{k}+i-2} - \bar{K}_{\bar{k}+i-1}\bar{C}A^{i-j}K\tilde{z}_{\bar{k}+j-1}.$$

It follows that

$$\bar{H}_{[n,m]}(i,j) = \bar{A}_{\bar{k}+i-1}\bar{H}_{[n,m]}(i-1,j) - \bar{K}_{\bar{k}+i-1}\bar{C}A^{i-j}K.$$

(*ii*) $\bar{G}_{[n,n]}(i,j)$ represents the transformation matrix form $w_{\bar{k}+j-1}$ to $\theta_{\bar{k}+i-1}$. Note that (6.14) implies $\bar{G}_{[n,n]}(i,j) = 0_n, \forall j \geq i$. When $j < i$, we have

$$\theta^w_{\bar{k}+i-1} = \bar{A}_{\bar{k}+i-1}\theta^w_{\bar{k}+i-2} + \bar{K}_{\bar{k}+i-1}\bar{C}A^{i-j-1}w_{\bar{k}+j-1}.$$

Then $\forall i \in [\![2,s+1]\!], j \in [\![1,i-1]\!]$, the recursion holds:

$$\bar{G}_{[n,n]}(i,j) = \bar{A}_{\bar{k}+i-1}\bar{G}_{[n,n]}(i-1,j) + \bar{K}_{\bar{k}+i-1}\bar{C}A^{i-j-1}.$$

(*iii*) Similarly, $\bar{M}$ and $\bar{E}$ can be obtained by

$$\bar{M}_{[n,m+\bar{m}]}(i,j) = \begin{cases} 0_{n\times(m+\bar{m})}, i < j \\ \bar{K}_{\bar{k}+i-1}, i = j \\ \bar{A}_{\bar{k}+i-1}\bar{M}_{[n,m+\bar{m}]}(i-1,j), i > j \end{cases}$$

$$\bar{E}_{[n,n]}(i,1) = \begin{cases} \bar{K}_{\bar{k}}\bar{C}, i = 1 \\ \bar{A}_{\bar{k}+i-1}\bar{E}_{[n,n]}(i-1,1) + \bar{K}_{\bar{k}+i-1}\bar{C}A^{i-1}, i > 1 \end{cases}$$

Remark 6.3. *The augmented vectors $\mathcal{E}$ and Θ have been expressed as linear functions of $\mathcal{Z}$, $\mathcal{W}$, $\mathcal{V}$, and e. The two equations in* (6.13) *and* (6.15) *serve as the basics for the subsequent design of optimal attacks. Note that all these coefficient matrices are independent of online data and thus can be calculated offline. In the case that only the original measurement is available ($\mathbb{I}_i = \mathbb{I}_{i-1} \cup \{y_i\}$), we have $\bar{K}_i = K, \bar{A}_i = (I_n - KC)A, \forall i \in [\![\bar{k},k]\!]$; the recursions for H, G, E and $\bar{H}, \bar{G}, \bar{M}, \bar{E}$ can be simplified.*

6.3.3 Optimal Attack Policy

We now design the attack coefficients in (6.10). In Theorem 6.1, it has been proved that the optimal attack $\tilde{z}_i$ is a linear function of θ_i. $\forall i \in [\![\bar{k}, k]\!]$, we consider a more general attack model:

$$\tilde{z}_i = \sum_{j=\bar{k}}^{i} F_j^i \theta_j + b_i,\ b_i \sim \mathcal{N}(0_m, \Phi_i). \tag{6.16}$$

Define the matrices

$$\begin{aligned} F &= \begin{bmatrix} F_{\bar{k}}^{\bar{k}} & 0_{m\times n} & \cdots & 0_{m\times n} \\ F_{\bar{k}}^{\bar{k}+1} & F_{\bar{k}+1}^{\bar{k}+1} & \cdots & 0_{m\times n} \\ \vdots & \vdots & \ddots & \vdots \\ F_{\bar{k}}^{k} & F_{\bar{k}+1}^{k} & \cdots & F_k^k \end{bmatrix},\ T = \begin{bmatrix} I_{ns} & 0_{ns\times n} \end{bmatrix}, \\ \mathcal{B} &= \begin{bmatrix} b_{\bar{k}}^{\mathrm{T}} & b_{\bar{k}+1}^{\mathrm{T}} & \cdots & b_k^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}},\ \Phi_{\mathcal{B}} = \mathbb{E}[\mathcal{B}\mathcal{B}^{\mathrm{T}}], \\ X &= \mathrm{blkdiag}\{Y_{\bar{k}}K, Y_{\bar{k}+1}K, \cdots, Y_k K\}, \end{aligned}$$

then (6.16) can be written in a compact form as

$$\mathcal{Z} = FT\Theta + \mathcal{B},\ \mathcal{B} \sim \mathcal{N}(0_{ms}, \Phi_{\mathcal{B}}), \tag{6.17}$$

the attack performance in (6.9) becomes

$$\hat{J}_{[\bar{k},k]} = \mathrm{Tr}\{T^{\mathrm{T}} X \mathbb{E}[\mathcal{Z}\mathcal{E}^{\mathrm{T}}]\}. \tag{6.18}$$

From (6.15) and (6.17), we have

$$\mathcal{Z} = FT\bar{H}\mathcal{Z} + FT\bar{G}\mathcal{W} + FT\bar{M}\mathcal{V} + FT\bar{E}e + \mathcal{B}. \tag{6.19}$$

It can be verified that $I_{ms} - FT\bar{H}$ is invertible. Define the new variable

$$\bar{F} = (I_{ms} - FT\bar{H})^{-1}F, \tag{6.20}$$

then (6.19) and (6.15) yield

$$\mathcal{Z} = \bar{F}T\bar{G}\mathcal{W} + \bar{F}T\bar{M}\mathcal{V} + \bar{F}T\bar{E}e + b, \tag{6.21}$$

$$\mathcal{E} = (H\bar{F}T\bar{G} + G)\mathcal{W} + (H\bar{F}T\bar{E} + E)e + H\bar{F}T\bar{M}\mathcal{V} + Hb, \tag{6.22}$$

where $b = (I_{ms} - FT\bar{H})^{-1}\mathcal{B}$ denotes the compensatory Gaussian noise. Define the constant matrices

$$\begin{aligned}
\hat{X} &= H^{\mathrm{T}}T^{\mathrm{T}}X,\\
\tilde{X} &= T\bar{G}\mathcal{P}_w G^{\mathrm{T}}T^{\mathrm{T}}X + T\bar{E}\mathcal{P}_e E^{\mathrm{T}}T^{\mathrm{T}}X,\\
\bar{X} &= T\bar{G}\mathcal{P}_w \bar{G}^{\mathrm{T}}T^{\mathrm{T}} + T\bar{E}\mathcal{P}_e \bar{E}^{\mathrm{T}}T^{\mathrm{T}} + T\bar{M}\mathcal{P}_v \bar{M}^{\mathrm{T}}T^{\mathrm{T}}.
\end{aligned}$$

From (6.21)–(6.22), the attack performance in (6.18) can be written as

$$\hat{J}_{[\bar{k},k]} = \mathrm{Tr}[(\bar{F}\bar{X}\bar{F}^{\mathrm{T}} + \Phi)\hat{X} + \bar{F}\tilde{X}], \tag{6.23}$$

where $\Phi \in \mathbb{S}_{+}^{ms}$ denote the covariance of b. The covariance of $\mathcal{Z}$ is obtained from (6.21), i.e.,

$$\mathbb{E}[\mathcal{Z}\mathcal{Z}^{\mathrm{T}}] = \bar{F}\bar{X}\bar{F}^{\mathrm{T}} + \Phi. \tag{6.24}$$

To satisfy the stealthiness constraint in $\mathbf{P}_{6.1}$, the diagonal blocks of $\mathbb{E}[\mathcal{Z}\mathcal{Z}^{\mathrm{T}}]$ should be constants that equal to Σ. Note that $F \in \mathbb{BL}_{[ms,ns]}(m,n)$; from (6.20) one can check that $\bar{F} \in \mathbb{BL}_{[ms,ns]}(m,n)$. It poses an additional constraint on $\bar{F}$.

Define an auxiliary variable $\bar{S} = \bar{F}\bar{X}\bar{F}^{\mathrm{T}} + \Phi$. Summarizing the above analysis, the optimal attack policy in (6.17) can be obtained by solving the following optimization problem:

$$\begin{aligned}
\mathbf{P}_{6.2}: \quad & \min_{\bar{F},\bar{S}} \quad \mathrm{Tr}(\bar{S}\hat{X} + \bar{F}\tilde{X})\\
\text{s.t. } & \bar{F}_{[m,n]}(i,j) = 0_{m\times n},\ \forall\ i \in [\![1, s-1]\!], j \in [\![i+1, s]\!],\\
& \bar{S}_{[m,m]}(i,i) = \Sigma,\ \forall i \in [\![1,s]\!],\\
& \bar{S} - \bar{F}\bar{X}\bar{F}^{\mathrm{T}} \succeq 0.
\end{aligned}$$

Then F^* is reconstructed from (6.20), i.e.,

$$F^* = \bar{F}^*(I_{ms} + T\bar{H}\bar{F}^*)^{-1}, \tag{6.25}$$

and the covariance of $\mathcal{B}$ is

$$\begin{aligned}
\Phi_{\mathcal{B}}^* &= (I_{ms} - F^*T\bar{H})\Phi^*(I_{ms} - F^*T\bar{H})^{\mathrm{T}}\\
&= (I_{ms} - F^*T\bar{H})(\bar{S}^* - \bar{F}^*\bar{X}\bar{F}^{*\mathrm{T}})(I_{ms} - F^*T\bar{H})^{\mathrm{T}}.
\end{aligned}$$

Remark 6.4. *$\boldsymbol{P}_{6.2}$ has linear objective and equality constraints; the last constraint can be reformulated as a linear matrix inequality by applying Schur complement. Thus the problem is convex and can be solved efficiently. Though the definition tells that $F \in \mathbb{BL}_{[ms,ns]}(m,n)$, but in numerical examples one can find that F^* in* (6.25) *is a block diagonal matrix. This in turn verifies the effectiveness of Theorem 6.1. Another difference from attacks with the maximum greedy performance is that $\Phi_{\mathcal{B}}^*$ is not necessarily block diagonal, indicating that the compensatory Gaussian noise can be mutually correlated.*

The estimation error covariances evolve according to

$$\tilde{P}_{k|k-1} = A\tilde{P}_{k-1|k-1}A^{\mathrm{T}} + Q, \tag{6.26}$$

$$\tilde{P}_{k|k} = \tilde{P}_{k|k-1} + K\Sigma K^{\mathrm{T}} - KF_k^k P_k^\theta - P_k^\theta (F_k^k)^{\mathrm{T}} K^{\mathrm{T}}, \tag{6.27}$$

with initial condition $\tilde{P}_{\bar{k}-1|\bar{k}-1} = (I_n - KC)\bar{P}$. At each instant, (6.26) is used in (6.8) to calculate P_k^θ, thus closing the loop of iteration (6.27).

6.3.4 General Cases

In $\mathbf{P}_{6.1}$, we assume that a single-step χ^2 detector is deployed to reveal anomalies. In this case, $\tilde{z}_k \sim \mathcal{N}(0_{m\times 1}, \Sigma)$ is the only stealthiness constraint; the compromised innovation $\{\tilde{z}_k\}$ can be correlated; the corresponding attack cannot completely bypass a multiple-step χ^2 detector. To tackle this issue, one can simply add another constraint to the optimization problem. $\mathbf{P}_{6.2}$ becomes

$$\begin{aligned}
\mathbf{P}_{6.3}: \quad & \min_{\bar{F},\bar{S}} \quad \mathrm{Tr}(\bar{S}\hat{X} + \bar{F}\tilde{X}) \\
\text{s.t. } & \bar{F}_{[m,n]}(i,j) = 0_{m\times n}, \ \forall\, i \in [\![1, s-1]\!], j \in [\![i+1, s]\!], \\
& \bar{S}_{[m,m]}(i,j) = 0_m, \ \forall i \in [\![1, s-1]\!], j \in [\![i+1, \tau_i]\!], \\
& \bar{S}_{[m,m]}(i,i) = \Sigma, \ \forall i \in [\![1, s]\!], \\
& \bar{S} - \bar{F}\bar{X}\bar{F}^{\mathrm{T}} \succeq 0.
\end{aligned}$$

where $\tau_i = \min\{i + \tau - 1, s\}$ and $\tau \in \mathbb{N}$ is the width of χ^2 detectors.

The corresponding attack policy leads to a sequence $\{\tilde{z}_k\}$ that is uncorrelated in a sliding window of size τ; thus it can completely deceive a τ-step χ^2

detector. Specifically, if the compromised innovation is required to be i.i.d. white Gaussian (the width of χ^2 detector is infinity), the second and third constraints reduce to

$$\bar{S} = \text{blkdiag}\underbrace{\{\Sigma, \cdots, \Sigma\}}_{k-\bar{k}+1 \text{ times}}.$$

This attack can completely bypass a whiteness detector. We also see that if a χ^2 detector with larger τ is deployed, there will be more constraints imposed on $\bar{S}$, leading to a greater optimal objective value; thus the worst-case attack performance can be reduced. This conclusion justifies the advantage of adopting multiple-step χ^2 detectors in practical systems. Compared with a single-step detector, a multiple-step one is more robust to disturbances and can also mitigate the attack impact by imposing a more strict stealthiness constraint.

6.4 Examples

In this section, we use numerical examples to illustrate the effectiveness of the proposed method. Let

$$Q = \begin{bmatrix} 1.2 & 0 \\ 0 & 6 \end{bmatrix}, R = \begin{bmatrix} 10 & 0 \\ 0 & 2 \end{bmatrix}.$$

The system parameters A and C are 2×2 matrices that are randomly generated with Matlab for 10,000 times. Let $\bar{k} = 31$ and $k = 34$; we compare the estimation performance loss within a 4-step interval.

Fig. 6.1 shows the difference between the interval performance and the summation of greedy performance in [98]. Both policies have made full utilization of online information; but the one in this work is designed to maximize $J_{[31,34]}$ directly. Since the performance difference is bounded below by zero, we claim that for all randomly generated systems, the attack policy in this chapter causes more severe estimation quality degradation in the considered horizon. Now consider a stable LTI process:

$$A = \begin{bmatrix} -0.0658 & 0 \\ 0.1972 & 0.9558 \end{bmatrix}, \; C = \begin{bmatrix} 2.9405 & 0.9765 \\ -1.9122 & -0.6873 \end{bmatrix}.$$

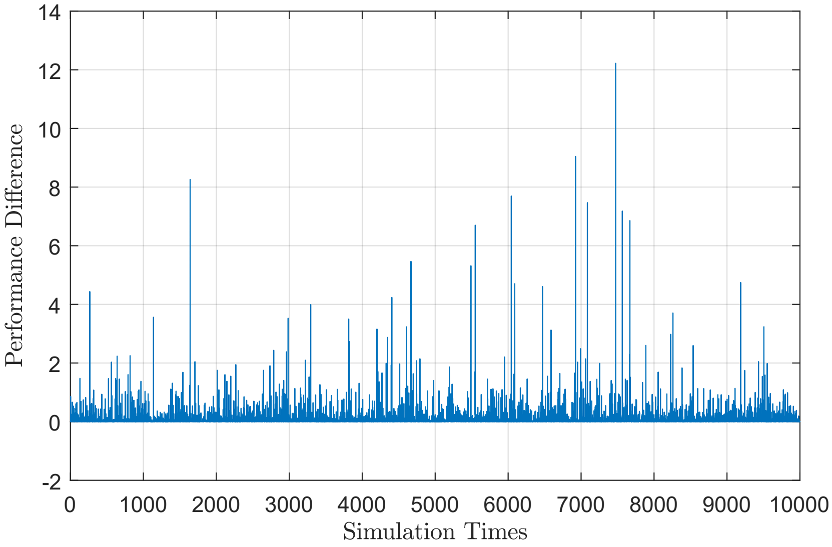

Figure 6.1: Difference of interval and greedy performance.

The interval performances with different attack policies are listed in Table 6.1. In [68], two historical innovations and the current one are utilized to design a linear attack. It is clear that the proposed attack in this work achieves the maximum interval performance.

Table 6.1: Attack performance comparison

Attack Policy	$\mathrm{Tr}(\tilde{P}_{31\vert 31} + \tilde{P}_{32\vert 32} + \tilde{P}_{33\vert 33} + \tilde{P}_{34\vert 34})$
[22]	229.5806
[68]	324.0134
[98]	342.0293
Theorem 6.1	342.0417

We then simulate the process for 10,000 times with randomly generated process and measurement noises. The threshold of χ^2 detector is set as 6, leading to a theoretical alarm rate 4.98%. Fig. 6.2 shows that the alarm rate remains unchanged after the attack occurs, thus verifying the stealthiness property of the proposed attack.

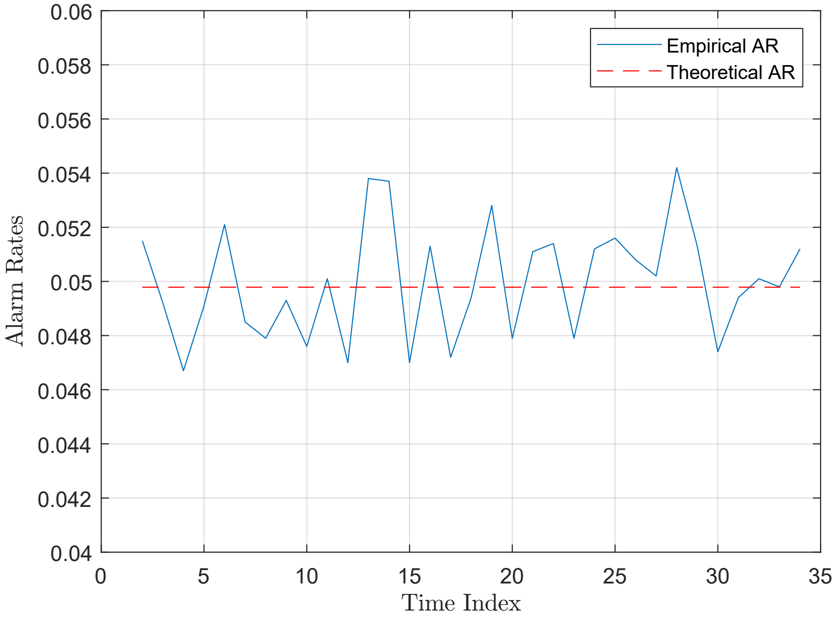

Figure 6.2: Theoretical and empirical alarm rates.

6.5 Conclusions

In this chapter, the optimal information-based attack that can maximize performance degradation of Kalman filters in a fixed interval and also completely bypass χ^2 detectors is studied. The attack is presupposed to be a linear combination of MMSE estimates of all historical innovations; then the combination coefficients can be obtained by solving a convex optimization problem. The proposed method can accommodate more general scenarios that an interval χ^2 detector is deployed to reveal anomalies; the worst-case attack impact is determined by the amount of online information and width of detector. Additionally, it can be shown that the optimal compromised measurements can be generated by an LPV system.

Chapter 7

Conclusions and Future Work

In this chapter, remarks are provided to conclude this thesis, and then some potential research directions are pointed out for future work.

7.1 Conclusions

This thesis focuses on the design of optimal cyber-attacks on industrial CPSs with energy and stealthiness constraints, aiming at revealing vulnerabilities of CPSs and establishing a basis for the development of defensive countermeasures. The outcomes of the studies in this thesis are summarized as follows:

1. We have studied a DoS attack problem against control channels in LQR systems, which has not been addressed in existing literature. Two common compensation strategies under DoS attacks are considered. Necessary and sufficient conditions are derived to ensure attack optimality from initial instants. A more general scenario that feasible attacks are not required to be consecutive is also briefly discussed.

2. We have proposed a novel linear attack policy based on the combined information and proved that it always outperforms the strategies using only partial information, which clarifies the counter-intuitive conclusion in [24]. More general scenarios are considered, including the correlated measurement noises between two sensors and time-varying means of the

injected bias. For attacks that can completely bypass χ^2 detectors, we give explicit solutions of the optimal attack policy, which avoids solving optimization problems numerically at each sampling instant. Uniqueness of the optimal stealthy attack strategy is analyzed. For scalar systems with uncorrelated measurement noises in two smart sensors, we give an easy-to-check criterion to compare the information fusion method with existing work.

3. We have revealed the connection between FDI attacks and MMSE estimation of prediction errors. The optimal stealthy attack policy that makes full utilization of attackers' available information and causes the maximum estimation performance loss is obtained. With theoretical analysis and comparative studies, it is proved that there does not exist another policy that outperforms the proposed one considering "greedy" performance criterion and the KL divergence stealthiness constraint. It is found that the optimal attack policy can be derived by recursively solving a constrained optimization problem, which is different but similar to LQG control. The well-known separation principle still holds; the optimal attack policy is a combination of MMSE estimation and linear transformation. For attacks with strict stealthiness, the optimal transformation matrix is obtained analytically without solving SDPs. For attacks with relaxed stealthiness, it is shown that the optimal compromised innovation is zero-mean Gaussian. The estimator with falsified measurements still provides unbiased state estimation. In contrast to existing literature that assumes static linear attacks, the optimal compromised measurement is shown to be generated by a linear time-varying (LTV) system, the coefficient matrices of which are independent of measurement data and thus can be determined offline. The cases that the attacker has different information sources are studied in a unified framework. It is found that the attack performance depends on the estimation quality for the prediction error. The attacker's additional information

will *always* benefit his/her purpose.

4. We have extended the information-based FDI attacks to the case that an interval anomaly detector is deployed to reveal anomalies. The optimal deception policy that makes full utilization of available online information and successfully deceives interval χ^2 detectors with different width is derived. Contrary to existing studies, the linearity assumption is removed and stealthiness constraints are fulfilled precisely. As a special case, if only the original measurement is available and the compromised innovation is required to be white Gaussian, the result reduces to the well-known conclusion in [22].

5. Finally, we have derived the optimal information-based FDI attack that maximizes the estimation error of Kalman filters in a fixed interval. It is shown that the information-based optimal attack is a linear combination of the MMSE estimates of all historical prediction errors. The combination coefficients can be obtained by solving a convex optimization problem. Moreover, the proposed attack can be generalized to deceive interval χ^2 detectors with different lengths by slightly modifying the stealthiness constraint. For both attacks with greedy and holistic performance, the worst-case attack impact is shown to be dependent on the amount of online information and the width of detection interval.

The findings in this thesis also provide some insights on defending against cyber-attacks in industrial CPSs. For stealthy FDI attacks producing compromised innovations with unaltered statistical properties, the detection mechanism based on only the probability density change of innovations is not sufficient to reveal anomalies. Though increasing the width of χ^2 detectors can mitigate the impact of worst-case attacks, defenders need to incorporate additional information or detection algorithms to completely resist stealthy attacks. In Fig. 3.1, it is not difficult to ensure $\tilde{z}_k \sim \mathcal{N}(0, \Sigma)$ if attackers can read and manipulate all measurements in the wireless channel. However, if the data packets of at least one sensor are secured by encryption mechanisms and cannot be modified, it will be much harder for attackers to design compromised innovations that have a known probability distribution. From the defenders' point of view, protecting only a subset of sensors can greatly enhance the security of remote state estimators; this can be achieved by either transmitting some measurements by wired channels or securing one dimension of y_k with data encryption. In addition to χ^2 detectors, monitoring the covariance change of state estimates directly can be utilized as a complementary detection strategy. These two detectors running in parallel may work more efficiently to resist stealthy attacks.

7.2 Future Work

The future research directions on CPS security are summarized with the following aspects:

1. Develop effective countermeasures that can mitigate the impacts of worst-case cyber-attacks. In this thesis, it is shown that for standard Kalman filters with Gaussian process and measurement noises, using statistical properties of innovation sequences to configure anomaly detectors cannot reveal the existence of stealthy FDI attacks efficiently. In order to resist FDI attacks that compromise remote state estimators, proactive defensive methods could be more suitable choices, such as encryption algorithms that prevent manipulation of transmitted data and physical authentication that enhances attack detectability with some sacrifice of control performance. Therefore, two research topics can be investigated: *i*) Designing optimal encryption algorithms that consume minimal computational resources and achieve maximum resistance effects; *ii*) Designing optimal water-marking signals that can prevent stealthy FDI attacks and have minimal or no negative impacts on the nominal system performance.

2. Explore worst-case deception attacks considering more general scenarios like transmission delays and packet dropouts, which are common in networked systems. A majority of existing work concentrates on only the case that transmission links are perfect when there is no cyber-attack. Little attention has been paid to these practical issues.

3. Study data-driven design methods for stealthy FDI attacks. In Section 5.1.4, three items are listed to describe attackers' ability. The second and third ones can be relaxed because even if attackers cannot gain any online information, they can still launch stealthy deception attacks if the transmitted raw data can be modified. The first assumption that attackers know all system parameters is the most important one in order to launch successful attacks. In practical cases, obtaining system parameters is not an easy task; thus studying the vulnerabilities of remote state estimation with limited knowledge is a meaningful topic. One possible research direction is investigating data-driven design methods that synthesize optimal attacks using only the online intercepted data.

Bibliography

[1] A. W. Al-Dabbagh, Y. Li, and T. Chen. An intrusion detection system for cyber attacks in wireless networked control systems. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 65(8):1049–1053, 2018.

[2] S. Amin, X. Litrico, S. Sastry, and A. M. Bayen. Cyber security of water SCADA systems–Part I: Analysis and experimentation of stealthy deception attacks. *IEEE Transactions on Control System Technology*, 21(5):1963–1970, 2013.

[3] B. D. O. Anderson and J. B. Moore. *Optimal Filtering.* Prentice Hall, 1979.

[4] C. Bai, V. Gupta, and F. Pasqualetti. On Kalman filtering with compromised sensors: Attack stealthiness and performance bounds. *IEEE Transactions on Automatic Control*, 62(12):6641–6648, 2017.

[5] C. Bai, F. Pasqualetti, and V. Gupta. Data-injection attacks in stochastic control systems: Detectability and performance tradeoffs. *Automatica*, 82:251–260, 2017.

[6] G. K. Befekadu, V. Gupta, and P. J. Antsaklis. Risk-sensitive control under Markov modulated denial-of-service (DoS) attack strategies. *IEEE Transactions on Automatic Control*, 60(12):3299–3304, 2015.

[7] R. Bhatia. *Positive Definite Matrices.* Princeton university press, 2009.

[8] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan. *Linear Matrix Inequalities in System and Control Theory.* SIAM, 1994.

[9] A. Chattopadhyay and U. Mitra. Security against false data injection attack in cyber-physical systems. *IEEE Transactions on Control of Network Systems*, 7(2):1015–1027, 2020.

[10] Y. Chen, S. Kar, and J. M. Moura. Cyber-physical attacks with control objectives. *IEEE Transactions on Automatic Control*, 63(5):1418–1425, 2017.

[11] J.-H. Cho, D. P. Sharma, H. Alavizadeh, S. Yoon, N. Ben-Asher, T. J. Moore, D. S. Kim, H. Lim, and F. F. Nelson. Toward proactive, adaptive defense: A survey on moving target defense. *IEEE Communications Surveys and Tutorials*, 22(1):709–745, 2020.

[12] T. M. Cover. *Elements of Information Theory.* John Wiley & Sons, 1999.

[13] N. Dey, A. S. Ashour, F. Shi, S. J. Fong, and J. M. R. Tavares. Medical cyber-physical systems: A survey. *Journal of Medical Systems*, 42(4):1–13, 2018.

[14] S. M. Dibaji, M. Pirani, D. B. Flamholz, A. M. Annaswamy, K. H. Johansson, and A. Chakrabortty. A systems and control perspective of CPS security. *Annual Reviews in Control*, 47:394–411, 2019.

[15] S. X. Ding. *Model-based fault diagnosis techniques: design schemes, algorithms, and tools.* Springer Science & Business Media, 2008.

[16] J. Falco, J. Falco, A. Wavering, and F. Proctor. *IT security for industrial control systems.* Citeseer, 2002.

[17] H. Fang, L. Xu, Y. Zou, X. Wang, and K.-K. R. Choo. Three-stage Stackelberg game for defending against full-duplex active eavesdropping attacks in cooperative communication. *IEEE Transactions on Vehicular Technology*, 67(11):10788–10799, 2018.

[18] M. S. Faughnan, B. J. Hourican, G. C. MacDonald, M. Srivastava, J.-P. A. Wright, Y. Y. Haimes, E. Andrijcic, Z. Guo, and J. C. White.

Risk analysis of unmanned aerial vehicle hijacking and methods of its detection. In *2013 IEEE Systems and Information Engineering Design Symposium*, pages 145–150, 2013.

[19] P. Freeman, R. Pandita, N. Srivastava, and G. J. Balas. Model-based and data-driven fault detection performance for a small UAV. *IEEE/ASME Transactions on Mechatronics*, 18(4):1300–1309, 2013.

[20] Z. Gao, C. Cecati, and S. X. Ding. A survey of fault diagnosis and fault-tolerant techniques—Part I: Fault diagnosis with model-based and signal-based approaches. *IEEE Transactions on Industrial Electronics*, 62(6):3757–3767, 2015.

[21] R. M. Gerdes, C. Winstead, and K. Heaslip. CPS: an efficiency-motivated attack against autonomous vehicular transportation. In *Proceedings of the 29th Annual Computer Security Applications Conference*, pages 99–108, 2013.

[22] Z. Guo, D. Shi, K. H. Johansson, and L. Shi. Optimal linear cyber-attack on remote state estimation. *IEEE Transactions on Control of Network Systems*, 4(1):4–13, 2017.

[23] Z. Guo, D. Shi, K. H. Johansson, and L. Shi. Worst-case stealthy innovation-based linear attack on remote state estimation. *Automatica*, 89:117–124, 2018.

[24] Z. Guo, D. Shi, K. H. Johansson, and L. Shi. Worst-case innovation-based integrity attacks with side information on remote state estimation. *IEEE Transactions on Control of Network Systems*, 6(1):48–59, 2019.

[25] T. M. Hoang, H. Q. Ngo, T. Q. Duong, H. D. Tuan, and A. Marshall. Cell-free massive mimo networks: Optimal power control against active eavesdropping. *IEEE Transactions on Communications*, 66(10):4724–4737, 2018.

[26] L. Hu, Z. Wang, Q. Han, and X. Liu. State estimation under false data injection attacks: Security analysis and system protection. *Automatica*, 87:176–183, 2018.

[27] A. Humayed, J. Lin, F. Li, and B. Luo. Cyber-physical systems security—a survey. *IEEE Internet of Things Journal*, 4(6):1802–1831, 2017.

[28] N. Hurley, Z. Cheng, and M. Zhang. Statistical attack detection. In *Proceedings of the 3rd ACM conference on Recommender systems*, pages 149–156, 2009.

[29] O. C. Imer, S. Yüksel, and T. Başar. Optimal control of LTI systems over unreliable communication links. *Automatica*, 42(9):1429–1439, 2006.

[30] A. Kanellopoulos and K. G. Vamvoudakis. A moving target defense control framework for cyber-physical systems. *IEEE Transactions on Automatic Control*, 65(3):1029–1043, 2019.

[31] S. Lai, B. Chen, T. Li, and L. Yu. Packet-based state feedback control under DoS attacks in cyber-physical systems. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 66(8):1421–1425, 2018.

[32] R. Langner. Stuxnet: Dissecting a cyberwarfare weapon. *IEEE Security and Privacy*, 9(3):49–51, 2011.

[33] T. Li, F. Tan, Q. Wang, L. Bu, J. Cao, and X. Liu. From offline toward real time: A hybrid systems model checking and CPS codesign approach for medical device plug-and-play collaborations. *IEEE Transactions on Parallel and Distributed Systems*, 25(3):642–652, 2013.

[34] Y. Li and T. Chen. Stochastic detector against linear deception attacks on remote state estimation. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 6291–6296. IEEE, 2016.

[35] Y. Li, D. E. Quevedo, S. Dey, and L. Shi. SINR-based DoS attack on remote state estimation: A game-theoretic approach. *IEEE Transactions on Control of Network Systems*, 4(3):632–642, 2016.

[36] Y. Li, L. Shi, and T. Chen. Detection against linear deception attacks on multi-sensor remote state estimation. *IEEE Transactions on Control of Network Systems*, 5(3):846–856, 2018.

[37] Y. Li, L. Shi, P. Cheng, J. Chen, and D. E. Quevedo. Jamming attacks on remote state estimation in cyber-physical systems: A game-theoretic approach. *IEEE Transactions on Automatic Control*, 60(10):2831–2836, 2015.

[38] Y. Li and G. Yang. Optimal stealthy false data injection attacks in cyber-physical systems. *Information Sciences*, 481:474–490, 2019.

[39] Y. Li and G. Yang. Optimal stealthy innovation-based attacks with historical data in cyber-physical systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(6):3401–3411, 2021.

[40] Y. Li and G. Yang. Optimal innovation-based deception attacks with side information against remote state estimation in cyber-physical systems. *Neurocomputing*, 2022.

[41] G. Liang, S. R. Weller, J. Zhao, F. Luo, and Z. Y. Dong. The 2015 Ukraine blackout: Implications for false data injection attacks. *IEEE Transactions on Power Systems*, 32(4):3317–3318, 2016.

[42] H. Liu, Y. Ni, L. Xie, and K. H. Johansson. An optimal linear attack strategy on remote state estimation. *IFAC-PapersOnLine*, 53(2):3527–3532, 2020.

[43] S. Liu, B. Chen, T. Zourntos, D. Kundur, and K. Butler-Purry. A coordinated multi-switch attack for cascading failures in smart grid. *IEEE Transactions on Smart Grid*, 5(3):1183–1195, 2014.

[44] Y. Liu, A. Liu, X. Liu, and M. Ma. A trust-based active detection for cyber-physical security in industrial environments. *IEEE Transactions on Industrial Informatics*, 15(12):6593–6603, 2019.

[45] Y. Liu, P. Ning, and M. K. Reiter. False data injection attacks against state estimation in electric power grids. *ACM Transactions on Information and System Security*, 14(1):1–33, 2011.

[46] L. Ljung. System identification. In *Signal analysis and prediction*, pages 163–173. Springer, 1998.

[47] Y. Z. Lun, A. D'Innocenzo, I. Malavolta, and M. D. Di Benedetto. Cyber-physical systems security: a systematic mapping study. *arXiv preprint arXiv:1605.09641*, 2016.

[48] B. Ly and R. Ly. Cybersecurity in unmanned aerial vehicles. *Journal of Cyber Security Technology*, 5(2):120–137, 2021.

[49] T. Macaulay and B. L. Singer. *Cybersecurity for industrial control systems: SCADA, DCS, PLC, HMI, and SIS*. CRC Press, 2011.

[50] F. Miao, Q. Zhu, M. Pajic, and G. J. Pappas. A hybrid stochastic game for secure control of cyber-physical systems. *Automatica*, 93:55–63, 2018.

[51] Y. Mo and B. Sinopoli. False data injection attacks in control systems. In *Proceedings of the 1st Workshop on Secure Control Systems*, pages 1–6, 2010.

[52] Y. Mo and B. Sinopoli. On the performance degradation of cyber-physical systems under stealthy integrity attacks. *IEEE Transactions on Automatic Control*, 61(9):2618–2624, 2015.

[53] Y. Mo, S. Weerakkody, and B. Sinopoli. Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs. *IEEE Control Systems Magazine*, 35(1):93–109, 2015.

[54] Y. Nakahira and Y. Mo. Attack-resilient H_2, H_∞, and L_1 state estimator. *IEEE Transactions on Automatic Control*, 63(12):4353–4360, 2018.

[55] J. K. Naufal, J. B. Camargo, L. F. Vismari, J. R. de Almeida, C. Molina, R. I. R. González, R. Inam, and E. Fersman. A^2CPS: A vehicle-centric safety conceptual framework for autonomous transport systems. *IEEE Transactions on Intelligent Transportation Systems*, 19(6):1925–1939, 2017.

[56] Y. Ni, Z. Guo, Y. Mo, and L. Shi. On the performance analysis of reset attack in cyber-physical systems. *IEEE Transactions on Automatic Control*, 65(1):419–425, 2019.

[57] M. Pajic, I. Lee, and G. J. Pappas. Attack-resilient state estimation for noisy dynamical systems. *IEEE Transactions on Control of Network Systems*, 4(1):82–92, 2016.

[58] M. Pajic, J. Weimer, N. Bezzo, P. Tabuada, O. Sokolsky, I. Lee, and G. J. Pappas. Robustness of attack-resilient state estimators. In *2014 ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS)*, pages 163–174. IEEE, 2014.

[59] G. Park, C. Lee, H. Shim, Y. Eun, and K. H. Johansson. Stealthy adversaries against uncertain cyber-physical systems: Threat of robust zero-dynamics attack. *IEEE Transactions on Automatic Control*, 64(12):4907–4919, 2019.

[60] F. Pasqualetti, F. Dörfler, and F. Bullo. Attack detection and identification in cyber-physical systems. *IEEE Transactions on Automatic Control*, 58(11):2715–2729, 2013.

[61] M. Porter, P. Hespanhol, A. Aswani, M. Johnson-Roberson, and R. Vasudevan. Detecting generalized replay attacks via time-varying dynamic watermarking. *IEEE Transactions on Automatic Control*, 66(8):3502–3517, 2020.

[62] J. Qin, M. Li, L. Shi, and X. Yu. Optimal denial-of-service attack scheduling with energy constraint over packet-dropping networks. *IEEE Transactions on Automatic Control*, 63(6):1648–1663, 2017.

[63] R. Romagnoli, S. Weerakkody, and B. Sinopoli. A model inversion based watermark for replay attack detection with output tracking. In *2019 American Control Conference (ACC)*, pages 384–390. IEEE, 2019.

[64] M. Sahebsara, T. Chen, and S. L. Shah. Optimal H-infinity filtering in networked control systems with multiple packet dropouts. *Systems and Control Letters*, 57(9):696–702, 2008.

[65] M. B. Salem, S. Hershkop, and S. J. Stolfo. A survey of insider attack detection research. *Insider Attack and Cyber Security*, pages 69–90, 2008.

[66] L. Schenato. To zero or to hold control inputs with lossy links? *IEEE Transactions on Automatic Control*, 54(5):1093–1099, 2009.

[67] J. Shang, M. Chen, and T. Chen. Optimal linear encryption against stealthy attacks on remote state estimation. *IEEE Transactions on Automatic Control*, 66(8):3592–3607, 2021.

[68] J. Shang and T. Chen. Optimal stealthy integrity attacks on remote state estimation: The maximum utilization of historical data. *Automatica*, 128:Article 109555, 2021.

[69] J. Shang, H. Yu, and T. Chen. Worst-case stealthy innovation-based linear attacks on remote state estimation under Kullback–Leibler divergence. *IEEE Transactions on Automatic Control*, Early Access, 2021.

[70] J. Shang, H. Yu, and T. Chen. Worst-case stealthy attacks on stochastic event-based state estimation. *IEEE Transactions on Automatic Control*, 2021, DOI:10.1109/TAC.2021.3071948.

[71] J. Shang, H. Yu, and T. Chen. Worst-case stealthy innovation-based linear attacks on remote state estimation under Kullback–Leibler divergence. *IEEE Transactions on Automatic Control*, 2021, DOI:10.1109/TAC.2021.3125430.

[72] J. Shang, J. Zhou, and T. Chen. Single-dimensional encryption against innovation-based stealthy attacks on remote state estimation. *Automatica*, 136:Article 110015, 2022.

[73] C. E. Shannon. Communication theory of secrecy systems. *The Bell System Technical Journal*, 28(4):656–715, 1949.

[74] T. Sui, Y. Mo, D. Marelli, X. Sun, and M. Fu. The vulnerability of cyber-physical system under stealthy attacks. *IEEE Transactions on Automatic Control*, 66(2):637–650, 2020.

[75] R. Taormina, S. Galelli, N. O. Tippenhauer, E. Salomons, and A. Ostfeld. Characterizing cyber-physical attacks on water distribution systems. *Journal of Water Resources Planning and Management*, 143(5):04017009, 2017.

[76] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson. A secure control framework for resource-limited adversaries. *Automatica*, 51:135–148, 2015.

[77] J. Tian, R. Tan, X. Guan, Z. Xu, and T. Liu. Moving target defense approach to detecting Stuxnet-like attacks. *IEEE Transactions on Smart Grid*, 11(1):291–300, 2019.

[78] D. Wang, J. Huang, Y. Tang, and F. Li. A watermarking strategy against linear deception attacks on remote state estimation under KL divergence. *IEEE Transactions on Industrial Informatics*, 17(5):3273–3281, 2020.

[79] Z. Wang, H. Song, D. W. Watkins, K. G. Ong, P. Xue, Q. Yang, and

X. Shi. Cyber-physical systems for water sustainability: challenges and opportunities. *IEEE Communications Magazine*, 53(5):216–222, 2015.

[80] Z. Wang, B. Zhao, and R. S. Blum. An overview of cybersecurity for natural gas networks: Attacks, attack assessment, and attack detection. *Security in Cyber-Physical Systems*, pages 255–285, 2021.

[81] S. Wankhede and D. Kshirsagar. DoS attack detection using machine learning and neural network. In *2018 4th International Conference on Computing Communication Control and Automation (ICCUBEA)*, pages 1–5. IEEE, 2018.

[82] G. Wu, J. Sun, and J. Chen. Optimal data injection attacks in cyber-physical systems. *IEEE Transactions on Cybernetics*, 48(12):3302–3312, 2018.

[83] J. Wu and T. Chen. Design of networked control systems with packet dropouts. *IEEE Transactions on Automatic Control*, 52(7):1314–1319, 2007.

[84] S. Wu, Z. Guo, D. Shi, K. H. Johansson, and L. Shi. Optimal innovation-based deception attack on remote state estimation. In *2017 American Control Conf. (ACC)*, pages 3017–3022. IEEE.

[85] Z. Wu and Q. He. Optimal switching sequence for switched linear systems. *SIAM Journal on Control and Optimization*, 58(2):1183–1206, 2020.

[86] D. Ye and T.-Y. Zhang. Summation detector for false data-injection attack in cyber-physical systems. *IEEE Transactions on Cybernetics*, 50(6):2338–2345, 2019.

[87] S. Z. Yong, M. Zhu, and E. Frazzoli. Switching and data injection attacks on stochastic cyber-physical systems: Modeling, resilient estimation, and attack mitigation. *ACM Transactions on Cyber-Physical Systems*, 2(2):1–2, 2018.

[88] J. Yu, L. Lu, Y. Chen, Y. Zhu, and L. Kong. An indirect eavesdropping attack of keystrokes on touch screen through acoustic sensing. *IEEE Transactions on Mobile Computing*, 20(2):337–351, 2019.

[89] H. Zhang, P. Cheng, L. Shi, and J. Chen. Optimal denial-of-service attack scheduling against linear quadratic Gaussian control. In *2014 American Control Conference*, pages 3996–4001. IEEE.

[90] J. Zhang, J. Sun, and H. Lin. Optimal DoS attack schedules on remote state estimation under multi-sensor round-robin protocol. *Automatica*, 127:109517, 2021.

[91] R. Zhang and P. Venkitasubramaniam. Stealthy control signal attacks in linear quadratic Gaussian control systems: Detectability reward tradeoff. *IEEE Transactions on Information Forensics and Security*, 12(7):1555–1570, 2017.

[92] T. Zhang and D. Ye. False data injection attacks with complete stealthiness in cyber–physical systems: A self-generated approach. *Automatica*, 120:109117, 2020.

[93] W. Zhang, M. S. Branicky, and S. M. Phillips. Stability of networked control systems. *IEEE Control Systems Magazine*, 21(1):84–99, 2001.

[94] Y. Zhang and J. Jiang. Bibliographical review on reconfigurable fault-tolerant control systems. *Annual Reviews in Control*, 32(2):229–252, 2008.

[95] Z. Zhang, P. Cheng, J. Wu, and J. Chen. Secure state estimation using hybrid homomorphic encryption scheme. *IEEE Transactions on Control Systems Technology*, 29(4):1704–1720, 2020.

[96] J. Zhou, W. Ding, and W. Yang. A secure encoding mechanism against deception attacks on multi-sensor remote state estimation. *IEEE Transactions on Information Forensics and Security*, 2022.

[97] J. Zhou, J. Shang, and T. Chen. Optimal linear FDI attacks with side information: A comparative study. In *4th IEEE International Conference on Industrial Cyber-Physical Systems*, pages 138–143, 2021.

[98] J. Zhou, J. Shang, and T. Chen. Worst-case stealthy false-data injection attacks on remote state estimation. In *47th Annual Conference of the IEEE Industrial Electronics Society*, pages 1–6, 2021.

[99] Y. Zhu and W. X. Zheng. Observer-based control for cyber-physical systems with periodic DoS attacks via a cyclic switching strategy. *IEEE Transactions on Automatic Control*, 65(8):3714–3721, 2020.

Appendix A

In this appendix, we collect proofs of some of the results in the thesis.

A.1 Proof of Proposition 3.2

For notation convenience, we use letters with lower cases to represent corresponding constant parameters ($a = A, k = K, w = W, c = C, \bar{c} = \bar{C}, \sigma = \Sigma, \theta = \Theta, p = P, \check{p} = \check{P}$) and $t \in \mathbb{N}$ to denote time index. First, consider the optimal attack in (3.14) at $t = \bar{t}$ (when attack starts). From (3.18)–(3.19), we obtain $P^{\alpha}_{\bar{t}} = p, P^{\beta}_{\bar{t}} = \theta$. In Theorem 3.2, it can be verified that $r_{\bar{t}} = m = 1$, then $\mathcal{W}_{\bar{t}}$ vanishes; it follows that

$$H^*_{\bar{t}} = -\sqrt{\sigma} V_c U_c^{\mathrm{T}} \Pi^{-\frac{1}{2}}, \;\; \Phi^*_{\bar{t}} = 0 \tag{A.1}$$

where V_c, U_c satisfy $U_c S_c V_c^{\mathrm{T}} = k\sqrt{\sigma}\Pi^{-\frac{1}{2}} \left[cp, \bar{c}\theta\right]^{\mathrm{T}}$. Note that this is a column vector; the SVD becomes

$$V_c = 1, S_c = \|k\sqrt{\sigma}\Pi^{-\frac{1}{2}} \left[cp, \bar{c}\theta\right]^{\mathrm{T}} \|_2.$$

It follows that

$$U_c = \frac{k\sqrt{\sigma}\Pi^{-\frac{1}{2}} \left[cp, \bar{c}\theta\right]^{\mathrm{T}}}{\|k\sqrt{\sigma}\Pi^{-\frac{1}{2}} \left[cp, \bar{c}\theta\right]^{\mathrm{T}} \|_2}. \tag{A.2}$$

The objective value in $\mathbf{P}_{3.2}$ becomes

$$\mathrm{Tr}(H^*_{\bar{t}} Y_{\bar{t}} K) = -\sqrt{\sigma} V_c U_c^{\mathrm{T}} \Pi^{-\frac{1}{2}} \begin{bmatrix} cp \\ \bar{c}\theta \end{bmatrix} k = -S_c.$$

Similarly, by Theorem 3.3, the objective value in (3.51) is

$$-\bar{S}_c = -\|k\sqrt{\sigma}\check{\Pi}^{-\frac{1}{2}}\left[c\check{p}, \bar{c}\check{p}\right]^{\mathrm{T}}\|_2.$$

In (3.14), we apply the constant attack strategy denoted by $\mathcal{A} : \{H_t^* = H_{\bar{t}}^*, \Phi_t^* = \Phi_{\bar{t}}^*\}, \forall t \geq \bar{t}$. Consider the recursions of P_t^α and P_t^β with policy $\mathcal{A}$. From (3.18), we have $P_{\bar{t}}^\alpha = P_{\bar{t}-1}^\alpha = p$ and

$$\begin{aligned} P_t^\alpha - P_{t-1}^\alpha &= a_1(P_{t-1}^\alpha - P_{t-2}^\alpha) + a_2(\bar{T}_{t-2} - \bar{T}_{t-1}) \\ &= \begin{cases} -a_1\theta\bar{c}k\bar{T}_a & t = \bar{t}+1 \\ a_1(P_{t-1}^\alpha - P_{t-2}^\alpha) & t \geq \bar{t}+2 \end{cases} \end{aligned} \tag{A.3}$$

where $a_1 = a^2(1-kc) > 0, a_2 = a_1\theta\bar{c}k$. $\bar{T}_a = H_{\bar{t}}^*[0;1]$, i.e., the second entry of $H_{\bar{t}}^*$. From (3.17), we have $\theta > 0$. According to [97], we obtain $\operatorname{sgn}(\bar{T}_a) = -\operatorname{sgn}(k\bar{c}\theta) = -\operatorname{sgn}(k\bar{c})$. It follows that $-a_1\theta\bar{c}k\bar{T}_a > 0$. Then (A.3) indicates $P_{t+1}^\alpha \geq P_t^\alpha, \forall t \in \mathbb{N}$. Following similar arguments, one can verify $P_{\bar{t}}^\beta = P_{\bar{t}-1}^\beta = \theta$ and $P_{t+1}^\beta \geq P_t^\beta, \forall t \in \mathbb{N}$. Both P_t^α and P_t^β are non-decreasing sequences with policy $\mathcal{A}$. Now we compare the objective values of $\mathcal{A}$ and $\{T_t^*, \bar{T}_t^*\}$ for $\mathbf{P}_{3.2}$ at instant t. Let $T_a = H_{\bar{t}}^*[1;0]$; we have

$$\begin{aligned} J_{opt}^t &= kcP_t^\alpha T_t^* + k\bar{c}P_t^\beta\bar{T}_t^* \overset{(a)}{\leq} kcP_t^\alpha T_a + k\bar{c}P_t^\beta\bar{T}_a \\ &= kcpT_a + k\bar{c}\theta\bar{T}_a + kc(P_t^\alpha - p)T_a + k\bar{c}(P_t^\beta - \theta)\bar{T}_a \\ &\overset{(b)}{\leq} kcpT_a + k\bar{c}\theta\bar{T}_a = J_{opt}^{\bar{t}} = -S_c \end{aligned} \tag{A.4}$$

where (a) is because that $\{T_t^*, \bar{T}_t^*\}$ is the optimal solution and (b) is from the facts that $\operatorname{sgn}(T_a) = -\operatorname{sgn}(kc), \operatorname{sgn}(\bar{T}_a) = -\operatorname{sgn}(k\bar{c})$ [97] and $P_t^\alpha \geq p, P_t^\beta \geq \theta$. The optimal objective values of $\mathbf{P}_{3.2}$ and (3.51) are denoted as J_{opt}^t and $\bar{J}_{opt}^t$, respectively. By Theorem 3.2, we have $\bar{J}_{opt}^t = -\bar{S}_c, \forall t \geq \bar{t}$.

If $Y^{\mathrm{T}}\Pi^{-1}Y > \check{Y}^{\mathrm{T}}\check{\Pi}^{-1}\check{Y}$, then $-S_c < -\bar{S}_c$. It follows from (A.4) that $J_{opt}^t < \bar{J}_{opt}^t, \forall t \in \mathbb{N}$. For scalar systems, this inequality is sufficient to show that the attack performance of (3.14) is better than that of (3.49). One the other hand, if (3.14) is a better policy, when attacks starts it holds that $-S_c < -\bar{S}_c$, which directly leads to $Y^{\mathrm{T}}\Pi^{-1}Y > \check{Y}^{\mathrm{T}}\check{\Pi}^{-1}\check{Y}$.

A.2 Proof of Proposition 4.1

We first show that (4.29) can be written as a linear combination of all nominal historical innovations. Notice that in case $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k\}$, we have $\check{y}_k = y_k, \check{v}_k = v_k, \tilde{C} = C, \tilde{R} = R$ and $K_k^{\xi} = K$, then the "virtual measurement" is

$$\check{z}_k = y_k - C\tilde{x}_{k|k-1} = z_k + C(\hat{x}_{k|k-1} - \tilde{x}_{k|k-1}). \tag{A.5}$$

From the dynamics of $\tilde{x}_{k|k-1}$ and $\hat{x}_{k|k-1}$:

$$\begin{aligned}\tilde{x}_{k|k-1} &= A\tilde{x}_{k-1|k-2} + AK\tilde{z}_{k-1}^{*}\\ \hat{x}_{k|k-1} &= A\hat{x}_{k-1|k-2} + AKz_{k-1}\end{aligned}$$

we have

$$\hat{x}_{k|k-1} - \tilde{x}_{k|k-1} = A(\hat{x}_{k|k-1} - \tilde{x}_{k|k-1}) + AK(z_{k-1} - \tilde{z}_{k-1}^{*}). \tag{A.6}$$

Define the state vector $\eta_k = \hat{x}_{k|k-1} - \tilde{x}_{k|k-1}$. Substituting (4.29) into (A.6), we obtain

$$\eta_k = A\eta_{k-1} - AKT_{k-1}^{*}\xi_{k-1} + AKz_{k-1} - AKb_{k-1}. \tag{A.7}$$

Define $F_k = (I - KC)(A - AKT_k^{*})$. Substituting (4.29) and (A.5) into (4.46), we have

$$\begin{aligned}\xi_k &= F_{k-1}\xi_{k-1} + K[z_k + C(\hat{x}_{k|k-1} - \tilde{x}_{k|k-1})] - (I - KC)AKb_{k-1}\\ &= F_{k-1}\xi_{k-1} + KC\eta_k + Kz_k - (I - KC)AKb_{k-1}.\end{aligned} \tag{A.8}$$

From (A.7)–(A.8), it can be obtained that

$$\begin{aligned}\xi_k =&(F_{k-1} - KCAKT_{k-1}^{*})\xi_{k-1} + KCA\eta_{k-1}\\ &+ L_k z_k + KCAKz_{k-1} - AKb_{k-1}.\end{aligned} \tag{A.9}$$

Now define the state variable $\theta_k \in \mathbb{R}^{2n}$ and matrices

$$\theta_k = \begin{bmatrix}\xi_k\\ \eta_k\end{bmatrix}, \bar{F}_{k-1} = \begin{bmatrix}A - KCA - AKT_{k-1}^{*} & KCA\\ -AKT_{k-1}^{*} & A\end{bmatrix}$$

$$\bar{L} = \begin{bmatrix} KCAK \\ AK \end{bmatrix}, \tilde{L} = \begin{bmatrix} K \\ 0 \end{bmatrix}, E = \begin{bmatrix} -AK \\ -AK \end{bmatrix}, \bar{T}_k = \begin{bmatrix} T_k^* & 0 \end{bmatrix}.$$

From (4.29), (A.7) and (A.9), we have

$$\theta_k = \bar{F}_{k-1}\theta_{k-1} + \bar{L}z_{k-1} + \tilde{L}z_k + Eb_{k-1} \tag{A.10}$$

$$\tilde{z}_k^* = \bar{T}_k\theta_k + b_k. \tag{A.11}$$

The compromised innovation $\tilde{z}_k^*$ in (4.29) is the output of the above LTV system. Notice that $\hat{x}_{\bar{k}|\bar{k}-1} = \tilde{x}_{\bar{k}|\bar{k}-1}$ leads to $\eta_{\bar{k}} = 0$; $\xi_{\bar{k}}$ is the MMSE estimate of $e_{\bar{k}|\bar{k}-1}$. From (A.5), we have

$$\tilde{z}_{\bar{k}} = z_{\bar{k}}.$$

Since $K_{\bar{k}}^{\xi} = K$, from (4.19), we obtain the initial value

$$\xi_{\bar{k}} = \bar{\xi}_{\bar{k}} + K_{\bar{k}}^{\xi}(\tilde{z}_{\bar{k}} - C\bar{\xi}_{\bar{k}}) = K_{\bar{k}}^{\xi}\tilde{z}_{\bar{k}} = Kz_{\bar{k}}.$$

Thus the initial condition of (A.10) is $\theta_{\bar{k}} = \left[(Kz_{\bar{k}})^{\mathrm{T}}, 0\right]^{\mathrm{T}}$. It is observed that the optimal attack policy in (A.11) is a linear combination of all nominal innovations in $[\![\bar{k}, k]\!]$, added by a compensatory Gaussian noise. The combination coefficients can be obtained recursively. For clarity, we write (A.11) as:

$$\tilde{z}_k^* = \sum_{i=\bar{k}}^{k} H_i z_i + \tilde{b}_k, \quad \tilde{b}_k \sim \mathcal{N}(0, \tilde{\Theta}) \tag{A.12}$$

where $H_i, i \in [\![\bar{k}, k]\!]$ and $\tilde{\Theta}$ are known parameters determined by Theorem 1. When $\tau_k = k - \bar{k}$, (52) is a presupposed innovation-based attack model:

$$\tilde{z}_k = \sum_{i=0}^{k-\bar{k}} T_k^{[i]} z_{k-i} + b_k, \quad b_k \sim \mathcal{N}\left(0, \Phi_k\right) \tag{A.13}$$

where $T_k^{[i]}$ and Φ_k are parameters to be optimized. With (A.13), the attack performance and stealthiness constraints in $\mathbf{P}_{4.2}$ have analytical forms; thus the optimal attacks can be obtained by solving the following optimization problem [68, Th. 1]:

$$\begin{aligned}
&\min_{T_k^{[i]},\Phi_k} - \operatorname{Tr}\left\{\sum_{i=0}^{\tau_k}\left[\bar{P}C^{\mathrm{T}}\Sigma^{-1}T_k^{[i]}\left(W_k^{[i]}\right)^{\mathrm{T}}\right]\right\} \\
&\text{s.t. } \sum_{i=0}^{\tau_k} T_k^{[i]}\Sigma\left(T_k^{[i]}\right)^{\mathrm{T}} + \Phi_k - \Sigma = 0, \\
&\qquad \Phi_k \succeq 0.
\end{aligned} \tag{A.14}$$

where $W_k^{[i]}$s are constant variables determined by system parameters. Note that when $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k\}$ and $\tau_k = k - \bar{k}$, [21] and this paper solve the same problem. The objective functions and stealthiness constraints in (A.14) and $\mathbf{P}_{4.2}$ are the same. We can study the equivalence of (A.12) and (A.13) by comparing their objective values.

Suppose the optimal attack obtained from (A.14) is $\tilde{z}_k^\star$. The objective values of $\tilde{z}_k^*$ for (A.14) and $\tilde{z}_k^*$ for $\mathbf{P}_{4.2}$ are denoted as $f_k(\tilde{z}_k^\star)$ and $f_k(\tilde{z}_k^\star)$, respectively. Since (A.12) is a special realization of (A.13) and also satisfies $\tilde{z}_k^* \sim \mathcal{N}(0, \Sigma)$, we know that $\tilde{z}_k^*$ is in the feasible region of (A.14). Because $\tilde{z}_k^\star$ is the optimal solution of (A.14), we have

$$f_k(\tilde{z}_k^*) \geq f_k(\tilde{z}_k^\star). \tag{A.15}$$

In the proof of Theorem 1, we have shown that $\tilde{z}_k^*$ is an optimal information-based strategy that is no worse than any other feasible attacks, which directly leads to

$$f_k(\tilde{z}_k^*) \leq f_k(\tilde{z}_k^\star). \tag{A.16}$$

The above two inequalities yield $f_k(\tilde{z}_k^*) = f_k(\tilde{z}_k^\star)$. In this sense, we claim that the optimal attack in (31) with $\mathbb{I}_k = \mathbb{I}_{k-1} \cup \{y_k\}$ and (52) with $\tau_k = k - \bar{k}$ are equivalent because they have the same attack performance.

If $r = m$, both the optimal attacks in (4.29) and (4.50) are unique. Because the involved optimization problem has a unique optimal solution, one can verify that these two attacks lead to the same $\tilde{z}_k$ and hence the same $\tilde{y}_k$.

A.3 Proof of Proposition 4.2

Denote the optimal solution to problem $\mathbf{P}_{4.5}$ as $\mathcal{O}_k^* = \{\mu_k^*, \Sigma_k^*, \hat{\Sigma}_k^*, T_k^*, \beta_k^*, \Theta_k^*\}$. At the $\bar{k}$th sampling instant, (4.19) implies $\hat{\mu}_{\bar{k}} = U^{\mathrm{T}}\mathbb{E}[\xi_{\bar{k}}] = 0$. The objective function becomes

$$f_{\bar{k}}(\hat{\Sigma}_{\bar{k}}, T_{\bar{k}}) = \mathrm{Tr}(-\hat{K}\hat{\Sigma}_{\bar{k}}\hat{K}^{\mathrm{T}} + 2\hat{K}T_{\bar{k}}\hat{P}_{\bar{k}}).$$

The corresponding optimization problem is

$$\begin{aligned} \min_{\mu_{\bar{k}}, \Sigma_{\bar{k}}, \hat{\Sigma}_{\bar{k}}, T_{\bar{k}}, \Theta_{\bar{k}}} \quad & f_{\bar{k}}(\hat{\Sigma}_{\bar{k}}, T_{\bar{k}}) \\ \text{s.t.} \quad & T_{\bar{k}}\hat{P}_{\bar{k}}T_{\bar{k}}^{\mathrm{T}} + \Theta_{\bar{k}} = \Sigma_{\bar{k}} && \text{(A.17)} \\ & \Sigma_{\bar{k}} + \mu_{\bar{k}}\mu_{\bar{k}}^{\mathrm{T}} = \hat{\Sigma}_{\bar{k}} && \text{(A.18)} \\ & g(\Sigma_{\bar{k}}, \hat{\Sigma}_{\bar{k}}) \leq 0 && \text{(A.19)} \\ & \Theta_{\bar{k}} \succeq 0. && \text{(A.20)} \end{aligned}$$

We first show (A.19) holds with equality for $\mathcal{O}_{\bar{k}}^*$. Assume $g(\Sigma_{\bar{k}}^*, \hat{\Sigma}_{\bar{k}}^*) = -\epsilon < 0$. There exist $\omega_{\bar{k}} \in \mathbb{R}^m$ and $\Omega_{\bar{k}} = \omega_{\bar{k}}\omega_{\bar{k}}^{\mathrm{T}}$ such that

$$\omega_{\bar{k}} \notin \ker(\hat{K}), \ \mathrm{Tr}(\Sigma^{-1}\Omega_{\bar{k}}) = \epsilon.$$

Let $\mu_{\bar{k}} \in \mathbb{R}^m$ satisfy $\mu_{\bar{k}}\mu_{\bar{k}}^{\mathrm{T}} = \mu_{\bar{k}}^*(\mu_{\bar{k}}^*)^{\mathrm{T}} + \Omega_{\bar{k}}$. By setting $\hat{\Sigma}_{\bar{k}} = \hat{\Sigma}_{\bar{k}}^* + \Omega_{\bar{k}}$, it can be verified that $\{\mu_{\bar{k}}, \Sigma_{\bar{k}}^*, \hat{\Sigma}_{\bar{k}}, T_{\bar{k}}^*, \Theta_{\bar{k}}^*\}$ fulfills the above constraints and

$$f_{\bar{k}}(\hat{\Sigma}_{\bar{k}}, T_{\bar{k}}^*) - f_{\bar{k}}(\hat{\Sigma}_{\bar{k}}^*, T_{\bar{k}}^*) = -\,\mathrm{Tr}(\hat{K}\Omega_{\bar{k}}\hat{K}^{\mathrm{T}}) < 0 \tag{A.21}$$

which contradicts the optimality of $\mathcal{O}_{\bar{k}}^*$. Therefore, we have $g(\Sigma_{\bar{k}}^*, \hat{\Sigma}_{\bar{k}}^*) = 0$.

Now fixing $\mu_{\bar{k}}^*, \Sigma_{\bar{k}}^*$ and $\hat{\Sigma}_{\bar{k}}^*$, we see that $\{T_{\bar{k}}^*, \Theta_{\bar{k}}^*\}$ is the solution to the optimization problem

$$\begin{aligned} \min_{T_{\bar{k}} \in \mathbb{R}^{m \times n}, \Theta_{\bar{k}} \in \mathbb{S}_+^m} \ & f_{\bar{k}}(\hat{\Sigma}_{\bar{k}}^*, T_{\bar{k}}) && \text{(A.22)} \\ \text{s.t.}\ & T_{\bar{k}}\hat{P}_{\bar{k}}T_{\bar{k}}^{\mathrm{T}} + \Theta_{\bar{k}} = \Sigma_{\bar{k}}^*. \end{aligned}$$

Denote $\bar{\Omega}_{\bar{k}} = \mu^*_{\bar{k}}(\mu^*_{\bar{k}})^{\mathrm{T}} \succeq 0$. Consider the candidate solution $\mathcal{O}^\star_{\bar{k}} = \{\mu^\star_{\bar{k}}, \Sigma^\star_{\bar{k}}, \hat{\Sigma}^\star_{\bar{k}}, T^\star_{\bar{k}}, \Theta^\star_{\bar{k}}\}$, where $\mu^\star_{\bar{k}} = 0, \Sigma^\star_{\bar{k}} = \Sigma^*_{\bar{k}} + \bar{\Omega}_{\bar{k}}$, $\hat{\Sigma}^\star_{\bar{k}} = \hat{\Sigma}^*_{\bar{k}}$, and $\{T^\star_{\bar{k}}, \Theta^\star_{\bar{k}}\}$ are obtained by solving

$$\begin{aligned} \min_{T_{\bar{k}} \in \mathbb{R}^{m\times n}, \Theta_{\bar{k}} \in \mathbb{S}^m_+} \quad & f_{\bar{k}}(\hat{\Sigma}^\star_{\bar{k}}, T_{\bar{k}}) \\ \text{s.t. } & T_{\bar{k}} \hat{P}_{\bar{k}} T^{\mathrm{T}}_{\bar{k}} + \Theta_{\bar{k}} = \Sigma^\star_{\bar{k}}. \end{aligned} \tag{A.23}$$

One can verify that $\mathcal{O}^\star_{\bar{k}}$ satisfies (A.17)–(A.18) and (A.20). Note that

$$\begin{aligned} g(\Sigma^\star_{\bar{k}}, \hat{\Sigma}^\star_{\bar{k}}) =& g(\Sigma^*_{\bar{k}}, \hat{\Sigma}^*_{\bar{k}}) + \ln|\Sigma^*_{\bar{k}}| - \ln|\Sigma^\star_{\bar{k}}| \\ =& \ln|\Sigma^*_{\bar{k}}| - \ln|\Sigma^*_{\bar{k}} + \bar{\Omega}_{\bar{k}}| < 0. \end{aligned} \tag{A.24}$$

This implies that $\mathcal{O}^\star_{\bar{k}}$ is a feasible solution. For (A.23), multiplying on both sides of the equality constraint by $(\Sigma^\star_{\bar{k}})^{-\frac{1}{2}}$, we have

$$(\Sigma^\star_{\bar{k}})^{-\frac{1}{2}} T_{\bar{k}} \hat{P}_{\bar{k}} T^{\mathrm{T}}_{\bar{k}} (\Sigma^\star_{\bar{k}})^{-\frac{1}{2}} + (\Sigma^\star_{\bar{k}})^{-\frac{1}{2}} \Theta_{\bar{k}} (\Sigma^\star_{\bar{k}})^{-\frac{1}{2}} \preceq I. \tag{A.25}$$

Denote $\hat{T}_{\bar{k}} = (\Sigma^\star_{\bar{k}})^{-\frac{1}{2}} T_{\bar{k}} \hat{P}^{\frac{1}{2}}_{\bar{k}}, Y_{\bar{k}} = \hat{P}^{\frac{1}{2}}_{\bar{k}} \hat{K} (\Sigma^\star_{\bar{k}})^{\frac{1}{2}}$. Note that the first term of $f_{\bar{k}}(\hat{\Sigma}^\star_{\bar{k}}, T_{\bar{k}})$ is a constant. (A.23) reduces to the following problem:

$$\min_{\hat{T}_{\bar{k}} \in \mathbb{R}^{m\times r}} \quad \mathrm{Tr}(\hat{T}_{\bar{k}} Y_{\bar{k}}) \;\text{ s.t. } \hat{T}_{\bar{k}} \hat{T}^{\mathrm{T}}_{\bar{k}} \preceq I.$$

The constraint implies that $\sigma_1(\hat{T}_{\bar{k}}) \leq 1$. The objective function satisfies

$$\begin{aligned} \mathrm{Tr}(\hat{T}_{\bar{k}} Y_{\bar{k}}) \geq & -\sum_{i=1}^m |\lambda_i(\hat{T}_{\bar{k}} Y_{\bar{k}})| \geq -\sum_{i=1}^m \sigma_i(\hat{T}_{\bar{k}} Y_{\bar{k}}) \\ \geq & -\sum_{i=1}^r \sigma_i(\hat{T}_{\bar{k}}) \sigma_i(Y_{\bar{k}}) \geq -\sum_{i=1}^r \sigma_i(Y_{\bar{k}}). \end{aligned} \tag{A.26}$$

Therefore, the minimal objective value for (A.23) is

$$\begin{aligned} f_{\bar{k}}(\hat{\Sigma}^\star_{\bar{k}}, T^\star_{\bar{k}}) &= -\hat{K} \hat{\Sigma}^\star_{\bar{k}} \hat{K}^{\mathrm{T}} - 2\sum_{i=1}^r \sigma_i(Y_{\bar{k}}) \\ &= -\hat{K} \hat{\Sigma}^\star_{\bar{k}} \hat{K}^{\mathrm{T}} - 2\,\mathrm{Tr}[(Y_{\bar{k}} Y^{\mathrm{T}}_{\bar{k}})^{\frac{1}{2}}]. \end{aligned} \tag{A.27}$$

Similarly, for (A.22), we have the minimal objective value

$$\begin{aligned} f_{\bar{k}}(\hat{\Sigma}^*_{\bar{k}}, T^*_{\bar{k}}) &= -\hat{K} \hat{\Sigma}^*_{\bar{k}} \hat{K}^{\mathrm{T}} - 2\sum_{i=1}^r \sigma_i(X_{\bar{k}}) \\ &= -\hat{K} \hat{\Sigma}^*_{\bar{k}} \hat{K}^{\mathrm{T}} - 2\,\mathrm{Tr}[(X_{\bar{k}} X^{\mathrm{T}}_{\bar{k}})^{\frac{1}{2}}] \end{aligned} \tag{A.28}$$

where $X_{\bar{k}} = \hat{P}_{\bar{k}}^{\frac{1}{2}} \hat{K} (\Sigma_{\bar{k}}^{*})^{\frac{1}{2}}$. It follows that

$$X_{\bar{k}} X_{\bar{k}}^{\mathrm{T}} - Y_{\bar{k}} Y_{\bar{k}}^{\mathrm{T}} = -\hat{P}_{\bar{k}}^{\frac{1}{2}} \hat{K} \bar{\Omega}_{\bar{k}} \hat{K}^{\mathrm{T}} \hat{P}_{\bar{k}}^{\frac{1}{2}} \preceq 0. \tag{A.29}$$

From (A.27)–(A.28), we have

$$\begin{aligned} f_{\bar{k}}(\hat{\Sigma}_{\bar{k}}^{\star}, T_{\bar{k}}^{\star}) - f_{\bar{k}}(\hat{\Sigma}_{\bar{k}}^{*}, T_{\bar{k}}^{*}) = 2 \sum_{i=1}^{r} [\sigma_i(X_{\bar{k}}) - \sigma_i(Y_{\bar{k}})] \\ =2 \operatorname{Tr}[(X_{\bar{k}} X_{\bar{k}}^{\mathrm{T}})^{\frac{1}{2}} - (Y_{\bar{k}} Y_{\bar{k}}^{\mathrm{T}})^{\frac{1}{2}}] \leq 0. \end{aligned} \tag{A.30}$$

The last inequality is derived from (A.29) and Lemma 4.2; the equality is attained when $\mu_{\bar{k}}^{*} \in \operatorname{Ker}(\hat{K})$. (A.30) implies that the objective value of $\mathcal{O}_{\bar{k}}^{\star}$ is no larger compared with that of $\mathcal{O}_{\bar{k}}^{*}$. If $\mu_{\bar{k}}^{*} \neq 0$, (A.24) is a strict inequality constraint. From (A.21), we see that there exists another feasible solution that is strictly better than $\mathcal{O}_{\bar{k}}^{\star}$. This contradicts the optimality of $\mathcal{O}_{\bar{k}}^{*}$. We have $\mu_{\bar{k}}^{*} = 0$. (4.54) leads to $\beta_{\bar{k}}^{*} = 0$. Recall that $\mathbf{P}_{4.5}$ should be solved at each sampling instant; from (4.57), we have $\hat{\mu}_{\bar{k}+1} = 0$. Following the same arguments it can be verified that $\beta_{k}^{*} = 0, \forall k \geq \bar{k}$.

If $r = m$, $Y_k = \hat{P}_{k}^{\frac{1}{2}} \hat{K} (\Sigma_{k}^{\star})^{\frac{1}{2}} \in \mathbb{R}^m$ is nonsingular. Then $\sigma_i(Y_k) \neq 0, \forall i \in [\![1, r]\!]$. In order to achieve the lower bound in (A.26), we have $\sigma_i(\hat{T}_k) = 1, \forall i \in [\![1, r]\!]$, i.e., $\hat{T}_k = I$. Then (A.25) implies $\Theta_k = 0$; hence, $b_k = 0, \forall k \geq \bar{k}$.

A.4 Proof of Theorem 4.3

Similar to the proof of Theorem 4.1, it can be shown that the optimal compromised innovation is a linear function of the MMSE estimate. Substituting (4.44) into $\mathbf{P}_{4.2}$, we have

$$\begin{aligned} \min_{T_k \in \mathbb{R}^{m \times n}, \Theta_k \in \mathbb{S}_{+}^{m}} \quad & \operatorname{Tr}(T_k P_k^{\xi} K) \\ \text{s.t. } & T_k P_k^{\xi} T_k^{\mathrm{T}} + \Theta_k = \Sigma. \end{aligned} \tag{A.31}$$

Let $\bar{T}_k = \Sigma^{-\frac{1}{2}} T_k \Psi_k \Lambda_k^{\frac{1}{2}}, \bar{K}_k = \Lambda_k^{\frac{1}{2}} \Psi_k^{\mathrm{T}} K \Sigma^{\frac{1}{2}}$; (A.31) is equivalent to the optimization problem

$$\min_{\bar{T}_k \in \mathbb{R}^{m \times n}} \quad \operatorname{Tr}(\bar{T}_k \bar{K}_k) \ \text{ s.t. } \bar{T}_k \bar{T}_k^{\mathrm{T}} - I \preceq 0. \tag{A.32}$$

The objective function satisfies

$$\mathrm{Tr}(\bar{T}_k\bar{K}_k) \geq -\sum_{i=1}^{\bar{r}} \sigma_i(\bar{T}_k)\sigma_i(\bar{K}_k) \geq -\sum_{i=1}^{\bar{r}} \sigma_i(\bar{K}_k). \tag{A.33}$$

Define the Lagrange function

$$\mathcal{L}(\bar{T}_k, \nu_k) = \mathrm{Tr}(\bar{T}_k\bar{K}_k) + \mathrm{Tr}[\nu_k(\bar{T}_k\bar{T}_k^{\mathrm{T}} - I)]$$

where $\nu_k \in \mathbb{R}^{m\times m}$ is the Lagrangian multiplier, which is symmetric owing to the symmetry of $\bar{T}_k\bar{T}_k^{\mathrm{T}} - I$. The stationary points satisfy $\nu_k \succeq 0$ and

$$\bar{K}_k^{\mathrm{T}} + 2\nu_k\bar{T}_k = 0 \tag{A.34}$$

$$\nu_k(\bar{T}_k\bar{T}_k^{\mathrm{T}} - I) = 0. \tag{A.35}$$

The above two equations lead to $\bar{K}_k^{\mathrm{T}}\bar{K}_k = 4\nu_k\bar{T}_k\bar{T}_k^{\mathrm{T}}\nu_k$ and $\nu_k\bar{T}_k\bar{T}_k^{\mathrm{T}}\nu_k = \nu_k^2$, respectively. Then $\bar{K}_k^{\mathrm{T}}\bar{K}_k = 4\nu_k^2$. It follows that

$$2\nu_k = (\bar{K}_k^{\mathrm{T}}\bar{K}_k)^{\frac{1}{2}}. \tag{A.36}$$

Because $\mathrm{Im}(\bar{K}_k^{\mathrm{T}}) = \mathrm{Im}[(\bar{K}_k^{\mathrm{T}}\bar{K}_k)^{\frac{1}{2}}]$, (A.34) is a consistent equation. The general solution is

$$\bar{T}_k^* = -\frac{1}{2}\nu_k^+\bar{K}_k^{\mathrm{T}} + (I - \nu_k^+\nu_k)\mathcal{W}_k$$

where $\mathcal{W}_k \in \mathbb{R}^{m\times s}$ is an arbitrary matrix. Substituting (A.36) into $\bar{T}_k^*$ yields

$$\bar{T}_k^* = -\Phi_k\Pi_k^{-\frac{1}{2}}\Phi_k^{\mathrm{T}}\bar{K}_k^{\mathrm{T}} + \bar{\Phi}_k\bar{\Phi}_k^{\mathrm{T}}\mathcal{W}_k. \tag{A.37}$$

It follows that

$$\begin{aligned}\bar{T}_k^*(\bar{T}_k^*)^{\mathrm{T}} = \bar{\Phi}_k\bar{\Phi}_k^{\mathrm{T}}\mathcal{W}_k\mathcal{W}_k^{\mathrm{T}}\bar{\Phi}_k\bar{\Phi}_k^{\mathrm{T}} - \Phi_k\Pi_k^{-\frac{1}{2}}\Phi_k^{\mathrm{T}}\bar{K}_k^{\mathrm{T}}\mathcal{W}_k^{\mathrm{T}}\bar{\Phi}_k\bar{\Phi}_k^{\mathrm{T}} \\ -\bar{\Phi}_k\bar{\Phi}_k^{\mathrm{T}}\mathcal{W}_k\bar{K}_k\Phi_k\Pi_k^{-\frac{1}{2}}\Phi_k^{\mathrm{T}} + \Phi_k\Phi_k^{\mathrm{T}}\end{aligned}$$

then the constraint in (A.32) becomes

$$\begin{bmatrix} I & -\Pi_k^{-\frac{1}{2}}\Phi_k^{\mathrm{T}}\bar{K}_k^{\mathrm{T}}\mathcal{W}_k^{\mathrm{T}}\bar{\Phi}_k \\ -\bar{\Phi}_k^{\mathrm{T}}\mathcal{W}_k\bar{K}_k\Phi_k\Pi_k^{-\frac{1}{2}} & \bar{\Phi}_k^{\mathrm{T}}\mathcal{W}_k\mathcal{W}_k^{\mathrm{T}}\bar{\Phi}_k \end{bmatrix} \preceq I.$$

It can be derived that

$$\bar{\Phi}_k^{\mathrm{T}} \mathcal{W}_k \mathcal{W}_k^{\mathrm{T}} \bar{\Phi}_k \preceq I,\ \bar{\Phi}_k^{\mathrm{T}} \mathcal{W}_k \bar{K}_k \Phi_k \Pi_k^{-\frac{1}{2}} = 0. \tag{A.38}$$

Let $\mathcal{X}_k = \bar{\Phi}^{\mathrm{T}} \mathcal{W}_k \in \mathbb{R}^{(m-\bar{r})\times s}$ be the free parameter, then (4.65) is obtained immediately from (A.38). Substituting $\mathcal{X}_k, \bar{K}_k$ into (A.37), we have the optimal solution

$$\bar{T}_k^* = -\Phi_k \Pi_k^{-\frac{1}{2}} \Phi_k^{\mathrm{T}} \Sigma^{\frac{1}{2}} K^{\mathrm{T}} \Psi_k \Lambda_k^{\frac{1}{2}} + \bar{\Phi}_k \mathcal{X}_k. \tag{A.39}$$

Since $\mathrm{Im}(\bar{\Phi}_k) = \mathrm{Ker}(\bar{K}_k^{\mathrm{T}} \bar{K}_k)$, $\mathrm{Ker}(\bar{K}_k) \subseteq \mathrm{Ker}(\bar{K}_k^{\mathrm{T}} \bar{K}_k)$, we have $\bar{K}_k \bar{\Phi}_k = 0$; then

$$\mathrm{Tr}(\bar{\Phi}_k \mathcal{X}_k \bar{K}_k) = \mathrm{Tr}(\mathcal{X}_k \bar{K}_k \bar{\Phi}_k) = 0.$$

Because the last term of $\bar{T}_k^*$ does not change the objective value, it follows that

$$\mathrm{Tr}(\bar{T}_k^* \bar{K}_k) = -\,\mathrm{Tr}(\Pi_k^{\frac{1}{2}}) = -\sum_{i=1}^{\bar{r}} \sigma_i(\bar{K}_k).$$

The lower bound in (A.33) is attained. Therefore, (A.39) gives the solution set for (A.32). Note that $T_k \Psi_k = \Sigma^{\frac{1}{2}} \bar{T}_k \Lambda_k^{-\frac{1}{2}}$; the solution to $\mathbf{P}_{4.3}$ satisfies

$$T_k^* = \Sigma^{\frac{1}{2}} \bar{T}_k^* \Lambda_k^{-\frac{1}{2}} \Psi_k^{\mathrm{T}} + \mathcal{Y}_k \bar{\Psi}_k^{\mathrm{T}} \tag{A.40}$$

$$\Theta_k^* = \Sigma^{\frac{1}{2}} [I - \bar{T}_k^* (\bar{T}_k^*)^{\mathrm{T}}] \Sigma^{\frac{1}{2}} \tag{A.41}$$

where $\mathcal{Y}_k \in \mathbb{R}^{m\times(n-s)}$ is an arbitrary matrix. The optimal parameters in (4.64) are obtained by substituting $\bar{T}_k^*$ into (A.40)–(A.41).

A.5 Proof of Proposition 4.3

From Theorem 4.1, we have

$$KT_{\bar{k}}^* = -KV\hat{\Sigma}^{\frac{1}{2}} V_{\bar{k}} U_{\bar{k}}^{\mathrm{T}} \hat{P}_{\bar{k}}^{-\frac{1}{2}} U^{\mathrm{T}}.$$

Substituting $K = U\hat{S}K^{\mathrm{T}}$ yields

$$\begin{aligned} KT_{\bar{k}}^* K &= -U\hat{S}\hat{\Sigma}^{\frac{1}{2}} V_{\bar{k}} U_{\bar{k}}^{\mathrm{T}} \hat{P}_{\bar{k}}^{-\frac{1}{2}} U^{\mathrm{T}} U \hat{S} V^{\mathrm{T}} \\ &= -U\hat{P}_{\bar{k}}^{-\frac{1}{2}} U_{\bar{k}} S_{\bar{k}} V_{\bar{k}}^{\mathrm{T}} V_{\bar{k}} U_{\bar{k}}^{\mathrm{T}} \hat{P}_{\bar{k}}^{-\frac{1}{2}} \hat{S} V^{\mathrm{T}} \\ &= -U\hat{P}_{\bar{k}}^{-\frac{1}{2}} (\hat{P}_{\bar{k}}^{\frac{1}{2}} \hat{S}\hat{\Sigma}\hat{S} \hat{P}_{\bar{k}}^{\frac{1}{2}})^{\frac{1}{2}} \hat{P}_{\bar{k}}^{-\frac{1}{2}} \hat{S} V^{\mathrm{T}}. \end{aligned} \tag{A.42}$$

Note that $\tilde{P}_{\bar{k}|\bar{k}-1} = \bar{P}, P^e_{\bar{k}} = (I - KC)\bar{P}$. From (4.25), we have $P^\xi_{\bar{k}} = K\Sigma K^\mathrm{T}$. Then $\hat{P}_{\bar{k}} = U^\mathrm{T} P^\xi_{\bar{k}} U = \hat{S}\hat{\Sigma}\hat{S}$. Substituting it into (A.42), we obtain

$$KT^*_{\bar{k}}K = -U\hat{S}V^\mathrm{T} = -K.$$

From (4.19), we have $\xi_{\bar{k}} = \mathbb{E}[\tilde{e}_{\bar{k}|\bar{k}-1}|\mathbb{I}_{\bar{k}}] = Kz_{\bar{k}}$. It follows that

$$K\tilde{z}^*_{\bar{k}} = K(T^*_{\bar{k}}\xi_{\bar{k}} + b_{\bar{k}}) = KT^*_{\bar{k}}Kz_{\bar{k}} + K\bar{V}\epsilon_{\bar{k}} = -Kz_{\bar{k}}.$$

If $\mathrm{rank}(C) = m$, K has full column rank. Then $\tilde{z}^*_{\bar{k}} = -z_{\bar{k}}$.

A.6 Proof of Proposition 5.1

We prove by induction. $\forall i < \bar{k}$, let $\hat{k} \in \mathbb{N}, \hat{k} \geq \bar{k}$, and assume the following statement holds:

$$\mathbb{E}[\tilde{z}_k z_i^\mathrm{T}] = 0_m, \forall k \in [\![\bar{k}, \hat{k}]\!]. \tag{A.43}$$

According to (5.17), (5.20) and (A.43), we have

$$\mathbb{E}[\tilde{z}_{\hat{k}+1} z_i^\mathrm{T}] = H^\phi_{\hat{k}+1} K^\mathrm{T} \mathbb{E}[\beta_{\hat{k}+1} z_i^\mathrm{T}], \tag{A.44}$$

where $H^\phi_{\hat{k}+1}$ is composed of the first m columns of $H_{\hat{k}+1}$. From (5.14a)–(5.14b), we have

$$\alpha_{\hat{k}} = A\alpha_{\hat{k}-1} - \bar{K}_{\hat{k}}\bar{C}\bar{\alpha}_{\hat{k}} + \bar{K}_{\hat{k}}\bar{C}x_{\hat{k}} + \bar{K}_{\hat{k}}\bar{v}_{\hat{k}}. \tag{A.45}$$

Combining (A.45) and (5.12) yields

$$\begin{aligned}\alpha_{\hat{k}} - \tilde{x}_{\hat{k}|\hat{k}-1} = {} & A(\alpha_{\hat{k}-1} - \tilde{x}_{\hat{k}-1|\hat{k}-2}) + \bar{K}_{\hat{k}}\bar{C}(x_{\hat{k}} - \bar{\alpha}_{\hat{k}}) \\ & + \bar{K}_{\hat{k}}\bar{v}_{\hat{k}} - AK\tilde{z}_{\hat{k}-1}.\end{aligned} \tag{A.46}$$

It follows that

$$\begin{aligned}\beta_{\hat{k}+1} = {} & A\beta_{\hat{k}} + \bar{K}_{\hat{k}+1}\bar{C}(x_{\hat{k}+1} - \bar{\alpha}_{\hat{k}+1}) \\ & + \bar{K}_{\hat{k}+1}\bar{v}_{\hat{k}+1} - AK\tilde{z}_{\hat{k}}.\end{aligned} \tag{A.47}$$

Since $x_{\hat{k}+1}-\bar{\alpha}_{\hat{k}+1}$ is independent of all historical measurements up to instant $\hat{k}$, it is easy to verify that $x_{\hat{k}+1}-\bar{\alpha}_{\hat{k}+1}$ and $\bar{v}_{\hat{k}+1}$ are independent of z_i; then (A.47) leads directly to $\mathbb{E}[\beta_{\hat{k}+1}z_i^{\mathrm{T}}] = A\mathbb{E}[\beta_{\hat{k}}z_i^{\mathrm{T}}]$. Repeating the same arguments yields

$$\mathbb{E}[\beta_{\hat{k}+1}z_i^{\mathrm{T}}] = A^{\hat{k}+1-\bar{k}}\mathbb{E}[\beta_{\bar{k}}z_i^{\mathrm{T}}]. \tag{A.48}$$

At the $\bar{k}$th instant, from (5.14a)–(5.14b), we obtain

$$\begin{aligned}\beta_{\bar{k}} &= \alpha_{\bar{k}} - \tilde{x}_{\bar{k}|\bar{k}-1} = \alpha_{\bar{k}} - \hat{x}_{\bar{k}|\bar{k}-1} = \alpha_{\bar{k}} - \bar{\alpha}_{\bar{k}} \\ &= \bar{K}_{\bar{k}}(\bar{y}_{\bar{k}} - \bar{C}\bar{\alpha}_{\bar{k}}) = \bar{K}_{\bar{k}}[\bar{C}(x_{\bar{k}} - \hat{x}_{\bar{k}|\bar{k}-1}) + \bar{v}_k]. \end{aligned} \tag{A.49}$$

Since z_i is determined by the set $\{\hat{x}_{0|-1}, y_0, \ldots, y_i\}$ while $x_{\bar{k}} - \hat{x}_{\bar{k}|\bar{k}-1}$ is independent of $\{\hat{x}_{0|-1}, y_0, \ldots, y_{\bar{k}-1}\}$, $\beta_{\bar{k}}$ is independent of z_i; i.e., $\mathbb{E}[\beta_{\bar{k}}z_i^{\mathrm{T}}] = 0_{n\times m}$. According to (A.44) and (A.48), we have $\mathbb{E}[\tilde{z}_{\hat{k}+1}z_i^{\mathrm{T}}] = 0_m$.

It is now sufficient to prove $\mathbb{E}[\tilde{z}_{\bar{k}}z_i^{\mathrm{T}}] = 0_m$. Note that $\tilde{z}_{\bar{k}} = H_{\bar{k}}\phi_{\bar{k}} + b_{\bar{k}}$; we have

$$\mathbb{E}[\tilde{z}_{\bar{k}}z_i^{\mathrm{T}}] = H_{\bar{k}}\mathbb{E}[\phi_{\bar{k}}z_i^{\mathrm{T}}] = H_{\bar{k}}K^{\mathrm{T}}\mathbb{E}[\beta_{\bar{k}}z_i^{\mathrm{T}}] = 0_m,$$

which completes the proof.

A.7 Proof of Theorem 5.5

The KL divergence constraint in $\mathbf{P}_{5.4}$ does not restrict that $\tilde{z}_k$ must have a specific form: neither be Gaussian distributed nor have zero mean. To prove Theorem 5.5, it is sufficient to show that $\forall k \geq \bar{k}$ the optimal compromised innovation has the following form:

$$\tilde{z}_k^* = H_k^*\theta_k + b_k,\ b_k \sim \mathcal{N}(0_{m\times 1}, \Phi_k^*), \tag{A.50}$$

where H_k^* and Φ_k^* are obtained by solving $\mathbf{P}_{5.4}$ after substituting $\tilde{z}_k = H_k\theta_k + b_k,\ b_k \sim \mathcal{N}(0_{m\times 1}, \Phi_k)$. We prove this statement by induction. Let $\hat{k} \in \mathbb{N}$ and $\hat{k} > \bar{k}$. Suppose that (A.50) holds $\forall k \in [\![\bar{k}, \hat{k}]\!]$. When $k = \hat{k}+1$, assume that an *arbitrary* feasible attack policy (not necessarily to be Gaussian or with zero mean) is given by

$$\tilde{z}_{\hat{k}+1} = \pi_{\hat{k}+1}(\mathbb{I}_{\hat{k}+1}), \tag{A.51}$$

then we define

$$\mu^{z}_{\hat{k}+1} = \mathbb{E}[\tilde{z}_{\hat{k}+1}],\ \Sigma^{z}_{\hat{k}+1} = \mathrm{Cov}[\tilde{z}_{\hat{k}+1}]. \tag{A.52}$$

Consider the following linear attack policy based on the MMSE estimate of $\hat{e}_k$:

$$\tilde{z}^{\star}_{\hat{k}+1} = H^{\star}_{\hat{k}+1}\theta_{\hat{k}+1} + b_{\hat{k}+1},\ b_{\hat{k}+1} \sim \mathcal{N}(\mu^{z}_{\hat{k}+1}, \Phi^{\star}_{\hat{k}+1}), \tag{A.53}$$

where $H^{\star}_{\hat{k}+1}$ and $\Phi^{\star}_{\hat{k}+1}$ are derived by solving

$$\begin{aligned} \min_{H_{\hat{k}+1},\Phi_{\hat{k}+1}} \quad & \mathrm{Tr}\left\{\begin{bmatrix} P^{\phi}_{\hat{k}+1} & \mathcal{M}_{\hat{k}+1}\end{bmatrix} H^{\mathrm{T}}_{\hat{k}+1}\right\} \\ \text{s.t.} \quad & H_{\hat{k}+1}\begin{bmatrix} P^{\phi}_{\hat{k}+1} & \mathcal{M}_{\hat{k}+1} \\ \mathcal{M}^{\mathrm{T}}_{\hat{k}+1} & \Sigma_{\hat{k}+1}\end{bmatrix} H^{\mathrm{T}}_{\hat{k}+1} + \Phi_{\hat{k}+1} = \Sigma^{z}_{\hat{k}+1}, \\ & H_{\hat{k}+1}\begin{bmatrix} \mathcal{M}_{\hat{k}+1} \\ \Sigma_{\hat{k}+1}\end{bmatrix} = 0_{m\times(\bar{\tau}_{\hat{k}+1}-m)}, && \text{(A.54)} \\ & \Phi_{\hat{k}+1} \succeq 0. && \text{(A.55)} \end{aligned}$$

The above optimization problem is obtained by substituting (A.53) into $\mathbf{P}_{5.4}$ and replacing the KL divergence constraint with the first equality constraint (restricting $\mathrm{Cov}[\tilde{z}_{\hat{k}+1}]$). Three facts are incorporated: *i*), $\theta_{\hat{k}+1}$ is independent of $b_{\hat{k}+1}$; *ii*), $\theta_{\hat{k}+1}$ has zero mean, which follows directly from (5.29) and the assumption that (A.50) holds $\forall k \in [\![\bar{k}, \hat{k}]\!]$; and *iii*), the second moment of $\tilde{z}^{\star}_{\hat{k}+1}$ is a constant and equals to that of $\tilde{z}_{\hat{k}+1}$ in (A.51), thus the first term of the objective function in $\mathbf{P}_{5.4}$ is omitted.

Note that the above optimization problem is similar to $\mathbf{P}_{5.2}$. Following the same arguments as in the proof of Theorem 5.1, one can verify that $\tilde{z}^{\star}_{\hat{k}+1}$ in (A.53) causes no greater objective value for $\mathbf{P}_{5.4}$ compared with $\tilde{z}_{\hat{k}+1}$ in (A.51). Additionally, since $\tilde{z}_{\hat{k}+1}$ and $\tilde{z}^{\star}_{\hat{k}+1}$ have the same covariance and $\tilde{z}^{\star}_{\hat{k}+1}$ is Gaussian, by using the fact that Gaussian distribution has the maximal differential entropy among all probability distributions with a specified covariance [12], it is easy to verify that $\tilde{z}^{\star}_{\hat{k}+1}$ satisfies the KL divergence constraint (see a similar proof in [23]). We conclude that for an arbitrary attack in (A.51), there *always* exists an MMSE estimate-based policy that causes no less attack performance and also satisfies the stealthiness constraint. It then suffices to limit the scope of our discussion to the linear attack based on MMSE estimate. Suppose that the optimal policy at instant $\hat{k}+1$ is given by

$$\tilde{z}^{\star}_{\hat{k}+1} = H^{\star}_{\hat{k}+1}\theta_{\hat{k}+1} + b_{\hat{k}+1},\ b_{\hat{k}+1} \sim \mathcal{N}(\mu^{\star}_{\hat{k}+1}, \Phi^{\star}_{\hat{k}+1}),$$

$H^{\star}_{\hat{k}+1}$, $\mu^{\star}_{\hat{k}+1}$ and $\Phi^{\star}_{\hat{k}+1}$ are obtained from the following optimization problem

$$\begin{aligned}\min_{\mathcal{A}_{\hat{k}+1}} \quad & \mathrm{Tr}(-K\tilde{\Sigma}_{\hat{k}+1}K^{\mathrm{T}}) + 2\left\{\begin{bmatrix}P^{\phi}_{\hat{k}+1} & \mathcal{M}_{\hat{k}+1}\end{bmatrix} H^{\mathrm{T}}_{\hat{k}+1}\right\} \\ \text{s.t.} \quad & H_{\hat{k}+1}\begin{bmatrix}P^{\phi}_{\hat{k}+1} & \mathcal{M}_{\hat{k}+1} \\ \mathcal{M}^{\mathrm{T}}_{\hat{k}+1} & \Sigma_{\hat{k}+1}\end{bmatrix} H^{\mathrm{T}}_{\hat{k}+1} + \Phi_{\hat{k}+1} = \hat{\Sigma}_{\hat{k}+1}, \\ & \hat{\Sigma}_{\hat{k}+1} + \mu_{\hat{k}+1}\mu^{\mathrm{T}}_{\hat{k}+1} = \tilde{\Sigma}_{\hat{k}+1}, \\ & g(\tilde{\Sigma}_{\hat{k}+1}, \hat{\Sigma}_{\hat{k}+1}) \leq 0, \\ & \text{(A.54) and (A.55).}\end{aligned}$$

where $\hat{\Sigma}_{\hat{k}+1}$ and $\tilde{\Sigma}_{\hat{k}+1}$ are the covariance and second moment of $\tilde{z}_{\hat{k}+1}$, respectively. Since $\tilde{z}^{\star}_{k}$ is Gaussian, $g(\tilde{\Sigma}_{\hat{k}+1}, \hat{\Sigma}_{\hat{k}+1})$ is the KL divergence constraint function:

$$g(\tilde{\Sigma}_{\hat{k}+1}, \hat{\Sigma}_{\hat{k}+1}) = \mathrm{Tr}(\Sigma^{-1}\tilde{\Sigma}_{\hat{k}+1}) + \ln\frac{|\Sigma|}{|\hat{\Sigma}_{\hat{k}+1}|} - m - 2\delta.$$

For brevity, we denote the optimal solution as $\mathcal{A}^{\star}_{\hat{k}+1} = \{H^{\star}_{\hat{k}+1}, \mu^{\star}_{\hat{k}+1}, \Phi^{\star}_{\hat{k}+1}, \hat{\Sigma}^{\star}_{\hat{k}+1}, \tilde{\Sigma}^{\star}_{\hat{k}+1}\}$. When $\mu^{\star}_{\hat{k}+1}$, $\hat{\Sigma}^{\star}_{\hat{k}+1}$ and $\tilde{\Sigma}^{\star}_{\hat{k}+1}$ are fixed, $\{H^{\star}_{\hat{k}+1}, \Phi^{\star}_{\hat{k}+1}\}$ is the solution of

$$\begin{aligned}\min_{H_{\hat{k}+1}, \Phi_{\hat{k}+1}} \quad & \mathrm{Tr}(-K\tilde{\Sigma}^{\star}_{\hat{k}+1}K^{\mathrm{T}}) + 2\left\{\begin{bmatrix}P^{\phi}_{\hat{k}+1} & \mathcal{M}_{\hat{k}+1}\end{bmatrix} H^{\mathrm{T}}_{\hat{k}+1}\right\} \\ \text{s.t.} \quad & H_{\hat{k}+1}\begin{bmatrix}P^{\phi}_{\hat{k}+1} & \mathcal{M}_{\hat{k}+1} \\ \mathcal{M}^{\mathrm{T}}_{\hat{k}+1} & \Sigma_{\hat{k}+1}\end{bmatrix} H^{\mathrm{T}}_{\hat{k}+1} + \Phi_{\hat{k}+1} = \hat{\Sigma}^{\star}_{\hat{k}+1}, \\ & \text{(A.54) and (A.55).}\end{aligned}$$

Assume $\mu^{\star}_{\hat{k}+1} \neq 0_{m\times 1}$. Consider the attack policy

$$\mathcal{A}^{*}_{\hat{k}+1} = \{H^{*}_{\hat{k}+1}, \mu^{*}_{\hat{k}+1}, \Phi^{*}_{\hat{k}+1}, \hat{\Sigma}^{*}_{\hat{k}+1}, \tilde{\Sigma}^{*}_{\hat{k}+1}\},$$

where $\mu^{*}_{\hat{k}+1} = 0_{m\times 1}$, $\hat{\Sigma}^{*}_{\hat{k}+1} = \tilde{\Sigma}^{*}_{\hat{k}+1} = \hat{\Sigma}^{\star}_{\hat{k}+1} + \mu^{\star}_{\hat{k}+1}(\mu^{\star}_{\hat{k}+1})^{\mathrm{T}}$, $H^{*}_{\hat{k}+1}$ and $\Phi^{*}_{\hat{k}+1}$ are obtained by solving

$$\begin{aligned}\min_{H_{\hat{k}+1},\Phi_{\hat{k}+1}} \quad & \mathrm{Tr}(-K\tilde{\Sigma}^{*}_{\hat{k}+1}K^{\mathrm{T}}) + 2\left\{\begin{bmatrix} P^{\phi}_{\hat{k}+1} & \mathcal{M}_{\hat{k}+1}\end{bmatrix} H^{\mathrm{T}}_{\hat{k}+1}\right\} \\ \text{s.t.} \quad & H_{\hat{k}+1}\begin{bmatrix} P^{\phi}_{\hat{k}+1} & \mathcal{M}_{\hat{k}+1} \\ \mathcal{M}^{\mathrm{T}}_{\hat{k}+1} & \Sigma_{\hat{k}+1}\end{bmatrix} H^{\mathrm{T}}_{\hat{k}+1} + \Phi_{\hat{k}+1} = \hat{\Sigma}^{*}_{\hat{k}+1}, \\ & \text{(A.54) and (A.55).}\end{aligned}$$

Note that $\hat{\Sigma}^{*}_{\hat{k}+1} \succeq \hat{\Sigma}^{\star}_{\hat{k}+1}, \tilde{\Sigma}^{*}_{\hat{k}+1} = \tilde{\Sigma}^{\star}_{\hat{k}+1}$. When $\tilde{\Sigma}_{\hat{k}+1}$ is fixed, $g(\tilde{\Sigma}_{\hat{k}+1}, \hat{\Sigma}_{\hat{k}+1})$ is a decreasing function with respective to $|\hat{\Sigma}_{\hat{k}+1}|$. It then can be verified that $\mathcal{A}^{*}_{\hat{k}+1}$ satisfies the KL stealthiness constraint and thus is a feasible attack policy. Denote the objective values of $\mathcal{A}^{\star}_{\hat{k}+1}$ and $\mathcal{A}^{*}_{\hat{k}+1}$ for the above two optimization problems as $f_{\hat{k}+1}(\mathcal{A}^{\star}_{\hat{k}+1})$ and $f_{\hat{k}+1}(\mathcal{A}^{*}_{\hat{k}+1})$, respectively. According to the explicit solution derived in Section 5.2.4 [see (5.42)], we have

$$\begin{aligned}& f_{\hat{k}+1}(\mathcal{A}^{*}_{\hat{k}+1}) - f_{\hat{k}+1}(\mathcal{A}^{\star}_{\hat{k}+1}) \\ =& 2\,\mathrm{Tr}[(\Omega^{\frac{1}{2}}_{\hat{k}+1}\hat{\Sigma}^{\star}_{\hat{k}+1}\Omega^{\frac{1}{2}}_{\hat{k}+1})^{\frac{1}{2}} - (\Omega^{\frac{1}{2}}_{\hat{k}+1}\hat{\Sigma}^{*}_{\hat{k}+1}\Omega^{\frac{1}{2}}_{\hat{k}+1})^{\frac{1}{2}}] \leq 0,\end{aligned}$$

where $\Omega_{\hat{k}+1} = P^{\phi}_{\hat{k}+1} - \mathcal{M}_{\hat{k}+1}\Sigma^{-1}_{\hat{k}+1}\mathcal{M}^{\mathrm{T}}_{\hat{k}+1}$. Additionally, it can be shown that the optimal attack policy must satisfy $\mathcal{D}_{\mathrm{KL}}(\tilde{z}^{\star}_{k}\|z_k) = \delta$. If $\mu^{\star}_{\hat{k}+1} \neq 0_{m\times 1}$, then $g(\tilde{\Sigma}^{*}_{\hat{k}+1}, \hat{\Sigma}^{*}_{\hat{k}+1}) < g(\tilde{\Sigma}^{\star}_{\hat{k}+1}, \hat{\Sigma}^{\star}_{\hat{k}+1}) = 0$. One can *always* find another attack policy that causes strictly smaller objective value of $\mathbf{P}_{5.4}$ compared with $\mathcal{A}^{*}_{\hat{k}+1}$. The above analysis contradicts the optimality of $\mathcal{A}^{\star}_{\hat{k}+1}$. It follows that $\mu^{\star}_{\hat{k}+1} = 0_{m\times 1}$.

When $k = \bar{k}$, we have $\tilde{z}^{\star}_{\bar{k}} = H^{\star}_{\bar{k}}\phi_{\bar{k}} + b_{\bar{k}}$. Following similar arguments one can verify that $\tilde{z}^{\star}_{\bar{k}}$ must have zero mean, i.e., $\mu^{\star}_{\bar{k}} = 0_{m\times 1}$. It follows that $\mu^{\star}_{k} = 0_{m\times 1}, \forall k \geq \bar{k}$. The optimal attack has the form in (A.50), which completes the proof. The recursion of $\mathcal{M}_k$ is re-defined similarly as in the analysis in Section 5.2.3. According to (5.30), the last term in (5.33) is replaced by $-AK\tilde{\Sigma}^{*}_{k}$. Note that Σ_k denotes the covariance of $\begin{bmatrix}\tilde{z}^{\mathrm{T}}_{k-1}, & \cdots, & \tilde{z}^{\mathrm{T}}_{k-\tau_k+1}\end{bmatrix}^{\mathrm{T}}$. Thus it is defined accordingly.

A.8 Proof of Lemma 6.1

Define the following variables:

$$\mathcal{X}_s = S^{\mathrm{T}}\mathcal{X},\ \hat{\mathcal{X}}_s = \mathbb{E}[\mathcal{X}_s|\mathbb{I}_x] = S^{\mathrm{T}}\hat{\mathcal{X}},\ \mathcal{P}_s = \mathbb{E}[\hat{\mathcal{X}}_s\hat{\mathcal{X}}_s^{\mathrm{T}}].$$

The proof is divided into two parts.

Part 1: We first assume that $\mathcal{Y}$ is a linear function of $\hat{\mathcal{X}}_s$ added by a Gaussian noise:

$$\mathcal{Y} = L\hat{\mathcal{X}}_s + b,\ b \sim \mathcal{N}(0_m, \Phi), \tag{A.56}$$

where $L \in \mathbb{R}^{m\times m}$ and $\Phi \in \mathbb{S}_+^m$ are parameters to be designed; b is independent of other variables. Substituting (A.56) into the objective and constraint yields the optimization problem:

$$\min_{L,\Phi}\ \mathrm{Tr}(L\mathcal{P}_s), \quad \text{s.t.}\ \ L\mathcal{P}_sL^{\mathrm{T}} + \Phi = \Sigma. \tag{A.57}$$

Note the constraint can be rewritten as $L\mathcal{P}_sL^{\mathrm{T}} - \Sigma \preceq 0$ to eliminate Φ. Define the Lagrange function

$$\mathcal{L}(\nu, L) = \mathrm{Tr}(L\mathcal{P}_s) + \mathrm{Tr}[\nu(L\mathcal{P}_sL^{\mathrm{T}} - \Sigma)],$$

where $\nu \in \mathbb{S}_m^+$ is the Lagrange multiplier. The station point satisfies $\nu \succeq 0$ and

$$\mathcal{P}_s + 2\nu L\mathcal{P}_s = 0, \tag{A.58}$$

$$L\mathcal{P}_sL^{\mathrm{T}} - \Sigma = 0. \tag{A.59}$$

Denote $\mathcal{Z} = \frac{1}{2}[(I_m - \mathcal{P}_s^{+}\mathcal{P}_s)\hat{\mathcal{Z}} - I_m]$; from (A.58), we have $L^{\mathrm{T}}\nu = \mathcal{Z}$, where $\hat{\mathcal{Z}} \in \mathbb{R}^{m\times m}$ is an arbitrary matrix. Multiplying on both sizes of (A.59) with ν and substituting $L^{\mathrm{T}}\nu$, we have $\mathcal{Z}^{\mathrm{T}}\mathcal{P}_s\mathcal{Z} = \nu\Sigma\nu$. It follows that

$$(\Sigma^{\frac{1}{2}}\nu\Sigma^{\frac{1}{2}})^2 = \Sigma^{\frac{1}{2}}\mathcal{Z}^{\mathrm{T}}\mathcal{P}_s\mathcal{Z}\Sigma^{\frac{1}{2}}.$$

Since $\mathcal{Z}^{\mathrm{T}}\mathcal{P}_s\mathcal{Z} = \frac{1}{4}\mathcal{P}_s$, the above equation yields

$$\nu = \frac{1}{2}\Sigma^{-\frac{1}{2}}(\Sigma^{\frac{1}{2}}\mathcal{P}_s\Sigma^{\frac{1}{2}})^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}.$$

Note that $\mathcal{P}_s$ can be singular. From $L^{\mathrm{T}}\nu = \mathcal{Z}$ we have $L = \nu^{+}\mathcal{Z}^{\mathrm{T}} + (I_m - \nu^{+}\nu)\bar{\mathcal{Z}}$, where $\bar{\mathcal{Z}} \in \mathbb{R}^{m\times m}$ is a matrix of free entries that satisfies $L\mathcal{P}_s L^{\mathrm{T}} \preceq \Sigma$. Assume $x \in \mathrm{Ker}(\nu)$, i.e., $\nu x = 0$; it can be verified that $\mathcal{P}_s x = 0$. Thus $\mathrm{Ker}(\nu) \subseteq \mathrm{Ker}(\mathcal{P}_s)$; we have $\mathcal{P}_s(I_m - \nu^{+}\nu) = 0$. The free parameter $\bar{\mathcal{Z}}$ does not affect the objective value in (A.57). Note that $\hat{\mathcal{Z}}$ can also be designed freely. We chose $\bar{\mathcal{Z}} = \hat{\mathcal{Z}} = 0_m$; then $\mathcal{Z} = -\frac{1}{2}I_m$, which yields

$$L^{*} = -\frac{1}{2}\nu^{+}, \tag{A.60}$$

$$\Phi^{*} = \Sigma - L^{*}\mathcal{P}_s(L^{*})^{\mathrm{T}}. \tag{A.61}$$

Part 2: We now prove that $\mathcal{Y}^{*}$ in (A.56) with L and Φ given in (A.60)–(A.61) is the vector that achieves the minimum objective compared with any feasible $\mathcal{Y}$.

Define $r = \mathrm{rank}(\nu)$ and the eigenvalue decomposition $\nu = \Psi\Pi\Psi^{\mathrm{T}}$, where $\Pi \in \mathbb{S}^{r}_{++}$, $\Psi \in \mathbb{R}^{m\times r}$ is an orthogonal matrix. Then $\nu^{+} = \Psi\Pi^{-1}\Psi^{\mathrm{T}}$. Since all choices of $\bar{\mathcal{Z}}$ and $\hat{\mathcal{Z}}$ yield the same objective value, we let $\hat{\mathcal{Z}} = 0_m$ and $\bar{\mathcal{Z}} = \lambda\mathcal{Z}^{\mathrm{T}} = -\frac{\lambda}{2}I_m$ where $\lambda > 0$ is a free parameter. It follows that

$$\begin{aligned} L^{*} &= -\frac{1}{2}[\nu^{+} + \lambda(I_m - \nu^{+}\nu)] \\ &= -\frac{1}{2}(\Psi\Pi^{-1}\Psi^{\mathrm{T}} + \lambda\bar{\Psi}\bar{\Psi}^{\mathrm{T}}) \\ &= -\frac{1}{2}\begin{bmatrix}\Psi & \bar{\Psi}\end{bmatrix}\begin{bmatrix}\Pi^{-1} & 0_{r\times(m-r)} \\ 0_{(m-r)\times r} & \lambda I_{m-r}\end{bmatrix}\begin{bmatrix}\Psi^{\mathrm{T}} \\ \bar{\Psi}^{\mathrm{T}}\end{bmatrix} \prec 0, \end{aligned}$$

where $\bar{\Psi} \in \mathbb{R}^{m\times(m-r)}$ is the orthogonal complement of Ψ. The covariance of b is $\Phi^{*} = \Sigma - L^{*}\mathcal{P}_s(L^{*})^{\mathrm{T}}$. For a given λ, Φ^{*} is a constant. Since b merely serves as a compensatory term to satisfy the constant in (A.57), it can be verified that $\mathcal{Y}^{*} = L^{*}\hat{\mathcal{X}}_s$ is the optimal vector that solves the following problem:

$$\min_{\mathcal{Y}} \ \mathrm{Tr}\{\mathbb{E}[\mathcal{Y}\mathcal{X}_s^{\mathrm{T}}]\}, \quad \text{s.t.} \ \ \mathcal{Y} \sim \mathcal{N}(0_m, \Sigma - \Phi^{*}). \tag{A.62}$$

Assume $\mathcal{Y}$ is an arbitrary vector that satisfies the constraint in (A.62). It is now sufficient to compare objective values of $\mathcal{Y}^{*}$ and $\mathcal{Y}$. Let $\mathcal{S}^{*} = \mathrm{Tr}\{\mathbb{E}[\mathcal{Y}^{*}\mathcal{X}_s^{\mathrm{T}}]\}$,

$\mathcal{S} = \mathrm{Tr}\{\mathbb{E}[\mathcal{Y}\mathcal{X}_s^{\mathrm{T}}]\}$. Since $\mathcal{Y}^* = L^*\hat{\mathcal{X}}_s$ is the MMSE estimate for $L^*\mathcal{X}_s$, we have

$$\begin{aligned}&\mathbb{E}[(\mathcal{Y}^* - L^*\mathcal{X}_s)(\mathcal{Y}^* - L^*\mathcal{X}_s)^{\mathrm{T}}]\\ &\preceq \mathbb{E}[(\mathcal{Y} - L^*\mathcal{X}_s)(\mathcal{Y} - L^*\mathcal{X}_s)^{\mathrm{T}}].\end{aligned}$$

It follows that

$$-(\mathcal{S}^* - \mathcal{S})(L^*)^{\mathrm{T}} - L^*(\mathcal{S}^* - \mathcal{S})^{\mathrm{T}} \preceq 0.$$

Since $-L^* \succ 0$, the inequality indicates $\mathrm{Tr}(\mathcal{S}^* - \mathcal{S}) \leq 0$. Thus $\mathcal{Y}^*$ achieves the minimum objective value compared with an arbitrary feasible vector. This completes the proof.

A.9 Proof of Lemma 6.2

According to (3.1) and (6.2)–(6.3), we have

$$\begin{aligned}\tilde{e}_{k|k} &= \tilde{e}_{k|k-1} - K\tilde{z}_k,\\ \tilde{e}_{k|k-1} &= A\tilde{e}_{k-1|k-1} + w_{k-1}.\end{aligned}$$

It follows from (4.26)–(4.27) that

$$\begin{aligned}\tilde{P}_{k|k} =& \mathbb{E}[\tilde{e}_{k|k}\tilde{e}_{k|k}^{\mathrm{T}}] = \mathbb{E}[(\tilde{e}_{k|k-1} - K\tilde{z}_k)(\tilde{e}_{k|k-1} - K\tilde{z}_k)^{\mathrm{T}}]\\ =& \mathbb{E}[\tilde{e}_{k|k-1}\tilde{e}_{k|k-1}^{\mathrm{T}}] + K\mathbb{E}[\tilde{z}_k\tilde{z}_k^{\mathrm{T}}]K^{\mathrm{T}} - K\mathbb{E}[\tilde{z}_k\tilde{e}_{k|k-1}^{\mathrm{T}}]\\ &- \mathbb{E}[\tilde{e}_{k|k-1}\tilde{z}_k^{\mathrm{T}}]K^{\mathrm{T}}\\ =& \mathbb{E}[(A\tilde{e}_{k-1|k-1} + w_{k-1})(A\tilde{e}_{k-1|k-1} + w_{k-1})^{\mathrm{T}}]\\ &+ K\Sigma K^{\mathrm{T}} - K\mathbb{E}[\tilde{z}_k\tilde{e}_{k|k-1}^{\mathrm{T}}] - \mathbb{E}[\tilde{e}_{k|k-1}\tilde{z}_k^{\mathrm{T}}]K^{\mathrm{T}}\\ =& A\tilde{P}_{k-1|k-1}A^{\mathrm{T}} + Q + K\Sigma K^{\mathrm{T}} - \mathbb{E}[\tilde{e}_{k|k-1}\tilde{z}_k^{\mathrm{T}}]K^{\mathrm{T}}\\ &- K\mathbb{E}[\tilde{z}_k\tilde{e}_{k|k-1}^{\mathrm{T}}]\end{aligned}$$

Iterating the above equation from $k = \bar{k}$ to k and using $\tilde{P}_{\bar{k}-1|\bar{k}-1} = (I_n - KC)\bar{P}$, one can verify that the holistic attack performance has the given form.