Anomaly detection using isomorphic analysis for false data injection attacks in industrial control systems

Anomaly Detection using Isomorphic Analysis for False Data Injection Attacks in Industrial Control Systems

Xinchen Zhang[a,b], Zhihan Jiang[a], Yulong Ding[b], Edith C.H. Ngai[a,*], Shuang-Hua Yang[b,c,*]

[a]*Department of Electric and Electronic Engineering, The University of Hong Kong,*
[b]*Shenzhen Key Laboratory of Safety and Security for Next Generation of Industrial Internet and Department of Computer Science and Engineering, Southern University of Science and Technology,*
[c]*Department of Computer Science, University of Reading,*

Abstract

As the Industrial Internet-of-Things (IIoT) evolves, a growing number of industrial control systems (ICSs) are connecting to the Internet, making them more vulnerable to malicious attacks. This paper addresses the detection of false data injection (FDI) attacks, a prevalent threat to open ICSs. We introduce an innovative anomaly detection technique using isomorphic analysis to safeguard ICSs against FDI attacks. Isomorphic analysis involves comparing transmitted signals with their expected values, which are derived from mathematical models or isomorphic components. For a comprehensive defense mechanism, we incorporate three specific detectors: the control signal detector, the actuating signal detector, and the sensor reading detector. Designed to detect FDI attacks across various parts of the ICS, these detectors ensure the integrity of all transmitted signals throughout the physical control system. While the control signal detector adopts a threshold method, the other two rely on statistical approaches. If an attack is detected, the detectors can correct tampered signals before they reach downstream components, enhancing the system's overall resilience and fault tolerance. The effectiveness of these detectors is supported by rigorous mathematical proofs. Moreover, our experimental findings further reveal the superiority of the isomorphic strategy over prior work in terms of detection rate, detection time delay, and system resilience.

Keywords: anomaly detection, false-data injection attacks, industrial control systems, secure control

1. Introduction

Industrial control systems (ICSs) refer to control systems used for industrial production. They play a significant role in critical national infrastructures, such as power grids, water treatment plants, natural gas, and refineries. Originally designed without intensive security considerations due to their isolated operation, modern ICSs have evolved into highly interconnected cyber-physical systems as the world becomes more digitally connected. Yet, this increased connectivity brings forth its own set of challenges. Recent years have witnessed an increasing trend in documented attacks on ICS infrastructures [1]. Compromises in ICSs, differing from traditional computer systems, have the potential to cause severe physical damages like cascading failures across national infrastructures. Hence, it becomes imperative to fortify ICSs against these threats.

[*] Co-corresponding author

Table 1: Mathematical models of attacks in ICSs [6].

Attack Type	Mathematical Model
DoS attacks	$\bar{s}[t] \in \emptyset$, when the DoS attack is successful at time t
FDI attacks	$\bar{s}[t] = s[t] + \Delta_s[t]$, where $\Delta_s[t]$ denotes the distortion of $s[t]$ caused by attackers
Replay attacks	$\bar{s}[t] \in S_t$, where S_t denotes the set of true signals accepted by attackers before time t

ICS-targeted attacks fall into three primary categories: denial-of-service (DoS) attacks, deception attacks, and direct attacks on physical processes or components [2]. DoS attacks flood the target with an overwhelming volume of spurious packets, overloading its network bandwidth [3]. This overload results in the compromised system's inability to deliver standard services or provide resource access [4], leading to easily noticeable functionality anomalies. By comparison, deception attacks that compromise data integrity are usually more subtle and stealthy. Typical deception attacks include replay attacks and false data injection (FDI) attacks. In replay attacks, attackers record and replay data from undisturbed periods during system disturbances, misleading operators and preventing them from taking essential corrective actions [5]. Conversely, FDI attacks tamper directly with system signals, leading to deviations from the expected patterns. Direct attacks can be viewed as an extension of FDI attacks, as malicious physical interference can yield similar system signal deviations. The mathematical models of DoS attacks and typical deception attacks are illustrated in Table 1, where $s[t]$ is the true signal not tampered with by attackers at time t and $\bar{s}[t]$ represents the received signal under attack at time t.

In this work, we introduce a novel system architecture equipped with three isomorphic detectors to detect FDI attacks in ICSs. These detectors perform isomorphic analysis, contrasting transmitted signals with their expected values, which are derived from homogeneous components or models. Notably, we adopt a broader understanding of FDI attacks: as long as any signal transmitted among controllers, actuators, and sensors deviates from the norm, it can be considered the outcome of an FDI attack, regardless of the underlying cause. Even if the intrusion originates in the IT network, but the impact has propagated to the control system, our system design also enables reliable attack detection of such intrusions.

A comparison between our methodology and existing research is presented in Table 2. Our approach is white-box and based on an explicit system model. In contrast to black-box methods, like those employing machine learning [7, 8, 9, 10], our white-box technique does not necessitate prior data collection. In situations requiring adaptability to environmental changes or user preferences, the white-box model offers a distinct advantage due to its thorough understanding of system dynamics. Among white-box detection techniques, the dynamic watermark detection method [11] is a representative one. This technique integrates an additional watermark into the original control system to detect anomalies caused by both FDI and replay attacks in sensor readings. However, introducing this external watermark can, to an extent, impair the system's overall performance, which is an undesirable outcome in practical applications. Moreover, its detection capability is primarily restricted to abnormal sensor readings. Different from the dynamic watermark detection approach [11], our technique mainly utilizes the system's inherent noise to detect FDI attacks, thereby eliminating the need for additional watermarks. Furthermore, we provide comprehensive protection, covering all signals within the control system, including control signals, actuating signals, and sensor readings. In contrast, the watermarking method mainly focuses on sensor readings. A distinctive feature of our methodology is its capacity to resist attacks by correcting malicious signals,

Table 2: Comparison of related work.

		No Pre-collected Data Needed	Adapts to Environment/ Preference Changes	FDI Attack Detection			Fault Torlerance	No Additional Noise Introduced
				Controller	Actuator	Sensor		
White-box	Isomorphic Analysis	✓	✓	✓	✓	✓	✓	✓
	Watermark	✓	✓	✗	✗	✓	✗	✗
Black-box (eg. machine learning)		✗	✗	✓	✓	✓	✓	✓

therefore endowing the system with a fault tolerance capability. Additionally, our system is compatible with dynamic watermarking [11] to detect replay attacks, enhancing its versatility.

The main contributions of our work are listed as follows:

- We introduce a novel system architecture that incorporates three isomorphic detectors, each applying isomorphic analysis—a method of comparing transmitted signals with expected values derived from mathematical models—to ensure comprehensive protection. By utilizing the inherent noise of the system, our approach not only detects anomalies but also corrects them across controllers, actuators, and sensors, achieving a level of all-encompassing defense not provided by existing methods.
- A rigorous mathematical proof is provided to establish the theoretical foundation for our method's ability and prove our method can theoretically detect any effective FDI attacks.
- Experiments are conducted to validate the effectiveness of our detection method. The results indicate our approach outperforms the conventional dynamic watermarking method in terms of detection rate, time delay, and minimized signal distortion during attacks.

The remainder of this paper is organized as follows. Section 2 presents a review of the related work. In Section 3, we clarify potential attack scenarios and discuss engineering solutions to guarantee the reliability of signals transmitted by isomorphic detectors. In Section 4, we elaborate on our isomorphic analysis detection method, providing the theoretical proofs of our method. Section 5 details the experimental setup, a comprehensive analysis of the compared performance of our proposed methods, and a case study demonstrating the effectiveness of our detection method. In Section 6, we evaluate resource consumption and discuss the limitations of our approach. Finally, Section 7 concludes our work and suggests directions for future research.

2. Related Work

Extensive research has been conducted on the anomaly detection of typical attacks on ICSs due to their critical roles in infrastructure. Anomaly detection methods for ICSs can be broadly categorized into model-based and data-driven approaches [5]. Model-based methods utilize explicit physical models, offering a 'white-box' perspective. In contrast, data-driven methods, often considered 'black-box' approaches, are advantageous in situations where a precise physical model is not available.

2.1. Data-driven Anomaly Detection

The rise of Machine Learning (ML), and notably Deep Learning (DL), has paved the way for advanced data-driven methods in recent years. These methodologies typically reconstruct

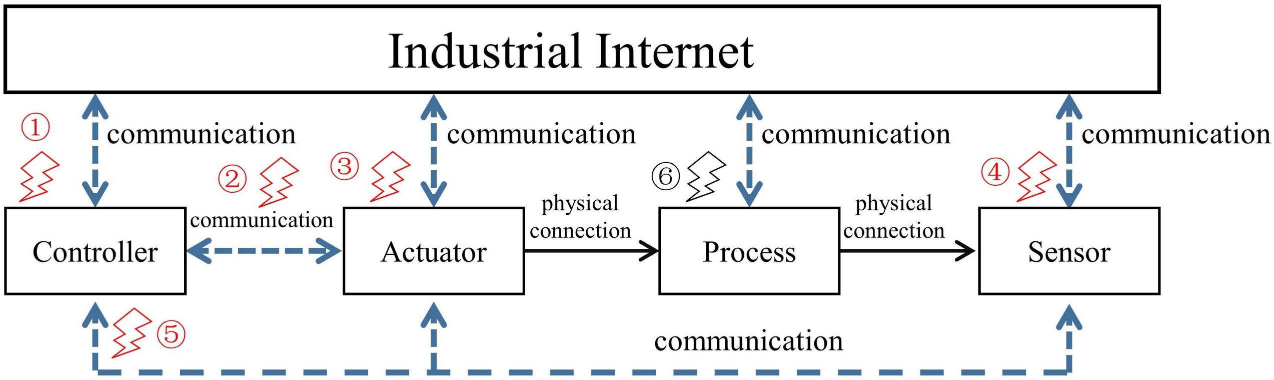


Figure 1: The architecture and vulnerabilities of the control system in ICS. Attacks in red numbers (attack ①-⑤) denote potential cyber-attacks. The attack (attack ⑥) in black numbers denotes potential direct attacks.

system signals based on learned patterns, using the reconstruction error as the primary criterion for anomaly detection [12]. Alternatively, some researchers approach anomaly detection as a classification task, aiming to distinguish the 'abnormal' class of signals using labeled datasets.

Prominent traditional ML techniques in ICSs include the support vector machine (SVM) [7], isolation forest (IF) [13], the Gaussian mixture model [8], and association rule mining [14]. Specifically, Esmalifalak, et al. [7] presented a machine learning methodology for identifying stealthy false data injections in smart grids. By leveraging supervised learning through a distributed SVM and employing Principal Component Analysis (PCA) for data reduction, their approach offers a robust solution for enhancing the security of power systems against sophisticated attacks. Ding et al. [13] developed an anomaly detection framework tailored for streaming data, employing the IF algorithm alongside a sliding window mechanism. This approach effectively addresses the dynamic nature of streaming data, demonstrating superior anomaly detection performance in real-time scenarios.

Deep learning models have further diversified the methods available. Examples include the long short-term memory (LSTM) network [9], the graph neural network (GNN) [15], and the transformer [10]. To be specific, Feng et al. [9] developed a multi-level anomaly detection framework for ICS, utilizing network package signatures alongside LSTM networks. Moreover, this approach combines regular communication patterns and a Bloom filter for efficient anomaly detection at the package content level, with a stacked LSTM network providing time-series level detection.

However, data-driven methods, while convenient due to their independence from explicit physical models, suffer from adaptability issues concerning the environment and user preferences. Their effectiveness is usually limited by the availability of a large number of training datasets.

2.2. Model-based Anomaly Detection

Model-based techniques utilize known physical models, providing insights grounded in system operational logic. Leveraging the explicit physical model of a control system, observers can be deployed to estimate the dynamics of a system [16, 17]. For instance, in [18], the authors proposed a real-time DoS attack detection scheme with a set of observers designed using sliding mode and adaptive estimation theory. As for FDI attacks, in [19], two adaptive sliding mode observers with online parameter estimation are designed to estimate state attacks and sensor attacks, respectively, showing that the constructed residual signals approach can detect the attacks with ultimately uniformly bounded errors. Manandhar et al. [20] adopted the Kalman filter to estimate the state process variables and proposed the χ^2-detector and Euclidean detector to build a robust security framework for the smart grid to detect system attacks, such as DoS attacks and FDI attacks. Similarly, Chen et al. [21] proposed extended state observer with H_∞ performance to estimate and detect FDI attacks for discrete-time nonlinear CPSs. Hu et al. [22] developed an FDI attack detection method for electric vehicle charging systems, utilizing time-frequency analysis to identify anomalies. This approach, while effective for its targeted FDI attacks on the charger or during the charger sensor signal transmission, cannot handle all kinds of FDI attacks, such as those targeted at the controller. Our methodology offers a more comprehensive FDI attack detection framework applicable across all components in ICSs.

A noteworthy technique is 'dynamic watermarking', proposed by Satchidanandan and Kumar [11]. They assessed the system's ability to respond appropriately to intentionally introduced signals, termed 'watermarks', by studying its closed-loop behavior in linear time-invariant systems with Gaussian noise models. Further development in this field led to the introduction of 'time-varying dynamic watermarking' as detailed in [23]. This method serves as an advanced linear time-varying adaptation of its predecessors, incorporating a matrix normalization factor specifically designed to address the system's temporal fluctuations. A foundational assumption of these dynamic watermarking techniques is the trustworthiness of the controllers and actuators, presupposing them to be uncompromised, ensuring the reliability of the transmitted data.

A critical observation from existing research, such as [24, 25, 20, 11, 21], is the focus on attacks compromising only measurements. This perspective often neglects the broader threat landscape, where controllers, actuators, and sensors are all potential targets. Recognizing this gap, our work proposes a novel architecture with isomorphic detection to comprehensively secure every physical component in the system.

3. Operational Feasibility

The interconnection of intelligent components to the Internet has transformed ICSs into open systems, making every networked component susceptible to cyber threats. To validate the realism and practicality of our application background, we detail potential attacks that an open ICS might confront and outline their execution. Additionally, we discuss engineering solutions to guarantee the reliability of signals transmitted by isomorphic detectors.

3.1. Attack Scenarios

Open ICSs, despite their improved control and communication capabilities, face increasing exposure to cyber threats. Figure 1 demonstrates potential direct and cyber-attacks on these systems, including direct attacks on the process (⑥) and cyber-attacks (①-⑤) on smart controllers, actuators, and sensors.

In Figure 1, attacks ①, ③, and ④ indicate scenarios where controllers, actuators, or sensors are compromised, leading to the generation of incorrect signals. Specifically, regarding the controller, attack ① can involve the manipulation of controller parameters. Concerning smart actuators, which possess embedded software or firmware that interprets control signals and converts them into appropriate actions, attack ③ can be updating this firmware with a

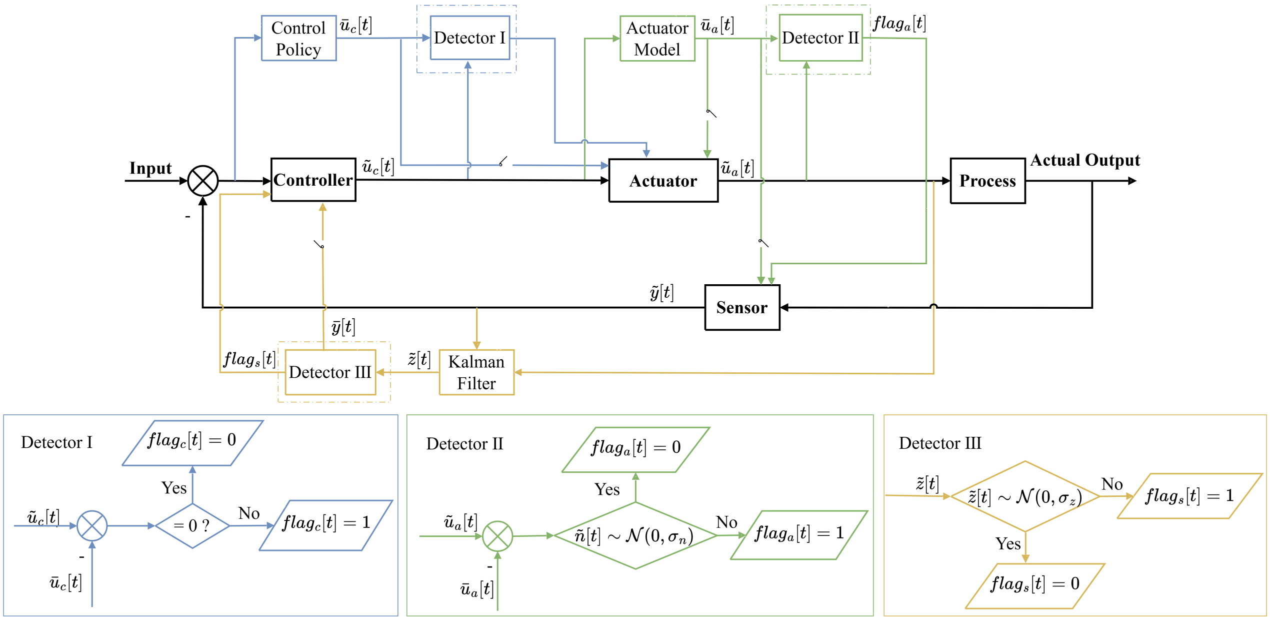


Figure 2: The architecture of the control system deployed with isomorphic detectors in ICS. The inner closed loop with bold black links denotes the real control system in operation. The blue, green, and yellow outer loops denote the logic of isomorphic detectors for the controller, actuator, and sensor, respectively.

malicious version to dictate how the actuator responds to legitimate control signals or even introduce new, unauthorized functionalities. As for the sensor, firmware is used to process the raw signals of some physical property (like temperature, pressure, light, etc.) for signal interpretation, compensation, calibration, etc. Attack ④ can be replacing legitimate firmware with a malicious version that sends false readings or allows for remote tampering.

Furthermore, attacks ② and ⑤ highlight the vulnerabilities during signal transmission for controllers and sensors, respectively. The action of the actuator is directly posed on the process, so there is no transmitted signal between the actuator and the process. Therefore, malicious alterations to actuating signals during transmission would only result in false alarms without affecting the physical systems. Hence, we do not consider the attack on the transmitted actuating signal.

On the other hand, attack ⑥ represents physical attacks on the process, such as contaminating a water supply or introducing impurities to a chemical process. While the rise of sophisticated cyber threats is undeniable, the age-old threat of physical tampering still exists. Protecting against both cyber and physical threats is vital for comprehensive security.

3.2. Reliability of Isomorphic Detectors

Figure 2 illustrates the role of isomorphic detectors in safeguarding the control system. Specifically, we utilize three isomorphic detectors, each designated for control signals, actuating signals, and sensor readings. These detectors identify anomalies by comparing transmitted signals to their expected values. If anomalies are detected in the transmitted signals, suggesting potential attacks, the isomorphic detectors employ two strategies to effectively counteract these disruptions. Firstly, when control signals or sensor readings are found malicious, they send the expected values to the downstream components. Secondly, in the presence of malicious actuating signals, they initiate corrective actions of the actuator. These expected values can be derived from mathematical models or isomorphic components. For example, to deduce the

Table 3: Summary of Signal and Function Notations.

Notation	**Description**
$g(\cdot)$	Control policy
$h(\cdot)$	Actuator modelling function
t	Time index
$\tilde{u}_c[t]$, $\tilde{u}_a[t]$, $\tilde{y}[t]$	Transmitted control signal, actuating signal, and measurement at time t, respectively
$u_c[t]$, $u_a[t]$, $y[t]$	True control signal, actuating signal, and measurement at time t, respectively
$\bar{u}_c[t]$, $\bar{u}_a[t]$, $\bar{y}[t]$	Expected control signal, actuating signal, and measurement at time t, respectively
$flag_c[t]$, $flag_a[t]$, $flag_s[t]$	Flags indicating malicious control signal, actuating signal, and measurement at time t, respectively
$n[t]$, $w[t]$, $v[t]$	Actuating noise, process noise and measurement noise at time t, respectively
$\Delta_a[t]$, $\Delta_n[t]$, $\Delta_z[t]$, $\Delta_x[t]$	Distortion of the actuating signal, actuating noise, innovation, and system state at time t, respectively
σ_n^2, σ_w^2, σ_v^2, σ_z^2	Variance of the actuating noise, process noise, measurement noise, and innovation, respectively
$\tilde{\mathbf{y}}[t]$	Historical measurements from the start to time $t-1$
$x[t]$, $\hat{x}^-[t]$, $\hat{x}[t]$, $z[t]$	True system state, prior and posterior estimates of the system state, and innovation at time t
$\tilde{n}[t]$, $\hat{\tilde{x}}^-[t]$, $\hat{\tilde{x}}[t]$, $\tilde{z}[t]$	Calculated actuating noise, prior and posterior estimates of the system state, and innovation at time t
$\mu[t]$	Self-adaptive coefficient for reliance on sensor readings at time t
$p[t]$	P-value of the statistical test used in detector III at time t
a, b, c, d	State coefficient, input coefficient, output coefficient, and direct transmission coefficient in the process model
k	Steady-state Kalman gain
P	Unique solution of the Riccati equation
K, S	Scale and shift hyperparameters in $\mu[t]$ function

expected value of a control signal, one can model the controller's control policy or employ an auxiliary controller with an identical control policy to mimic the behavior of the functioning controller. The expected values of transmitted signals provide a reference for anticipated system behaviors. Preserving their integrity is paramount to the reliability of our detection methodology.

There are many protective measures in engineering to achieve this. To name a few: 1) Network segmentation: keep the isomorphic components detached from the main network and operate them on a secured, exclusive server, sheltering the virtual components from potentially vulnerable parts of the original network. 2) Immutable infrastructure: establish a structure where virtual components, once set up, remain unchanged. Any changes or updates are made by replacing the entire system rather than editing the existing one. This limits the ability of an attacker to make persistent changes. 3) Access control: tighten access, allowing only a select few trusted individuals to modify the virtual component setup, or employ multi-factor authentication to enhance security.

Incorporating these strategies ensures the expected values remain a reliable metric for anomaly detection, reinforcing the trustworthiness of our isomorphic detectors.

4. Anomaly Detection using Isomorphic Analysis for False Data Injection Attacks

Consider a single-output-single-input (SISO) control system, which typically includes a controller, an actuator, a sensor, and a physical process as the controlled object. For instance, a water level control system is a common example of such a system in industrial settings, where maintaining a critical parameter—like the water level—at a specific setpoint is essential. In this system, the controller determines the desired water level, a valve acts as an actuator to adjust the water flow, a sensor monitors the current water level, and the tank itself is the physical process.

In this context, we propose a general architecture with isomorphic detectors to safeguard signals transmitted throughout the system and strive to maintain the normal operation of the system even in the face of FDI attacks. In Figure 2, the bold black inner loop represents the control system in operation, and the blue, green, and yellow external loops denote the logic of isomorphic detectors designated for the control signals, actuating signals, and sensor readings, respectively. The isomorphic detectors are deployed to verify the consistency of control signals, actuating signals, and sensor readings. They validate received data based on the isomorphic components or models.

Notations of Signals. As detailed in Table 3, $\tilde{u}_c[t]$, $\tilde{u}_a[t]$, and $\tilde{y}[t]$ denote the control signal, the actuating signal, and the measurement transmitted in the system at time t. Here, $t \in \mathbb{N}$ indicates the time index. $u_c[t]$, $u_a[t]$, and $y[t]$ denote the true and honest control signal, actuating signal, and measurement, which are only known to the attacker. The expected values for the control, actuating, and sensor signals at time t, derived from the control policy, actuator model, system model, and Kalman filter, are represented by $\bar{u}_c[t]$, $\bar{u}_a[t]$, and $\bar{y}[t]$. $flag_c[t]$, $flag_a[t]$, and $flag_s[t]$ indicate whether the control signal, actuating signal, and sensor readings are malicious or not at time t.

4.1. Detector I for the Control Signal

In Figure 1, attacks ① and ② are targeted for control signals. The process for testing control signals includes the calculation of expected control signal $\bar{u}_c[t]$ and the test of consistency between transmitted control signal $\tilde{u}_c[t]$ and expected control signal $\bar{u}_c[t]$. The following control policy is used to calculate true control signal $u_c[t]$.

Control Policy. Consider the controller using a history-dependent control policy to generate control signals. The expected value of the control signal at time t can be calculated by

$$\bar{u}_c[t] = g(\tilde{\mathbf{y}}[t-1]), \tag{1}$$

where $g(\cdot)$ denotes the control policy; $\tilde{\mathbf{y}}[t] := [\tilde{y}[0], \tilde{y}[1], ..., \tilde{y}[t-1]]$ denotes all the historical measurements. The control policy can be deployed in the downstream smart actuator through the program or conducted on a redundant controller operating under the same control law.

Test for Control Signals. Check whether the difference between transmitted control signal $\tilde{u}_c[t]$ and expected control signal $\bar{u}_c[t]$ satisfies:

$$\tilde{u}_c[t] - \bar{u}_c[t] = 0. \tag{2}$$

Since the control signals are electrical signals and the control policy is accurate, the difference between transmitted control signal $\tilde{u}_c[t]$ and expected control signal $\bar{u}_c[t]$ should be strictly zero.

Fault Tolerance. If transmitted control signal $\tilde{u}_c[t]$ and expected control signal $\bar{u}_c[t]$ pass the test, the flag indicating malicious control signals $flag_c[t]$ should be 0, which means the controller is honest and reports the correct control signal, and vice versa. If $flag_c[t]$ is 1, the downstream actuator should execute expected control signal $\bar{u}_c[t]$ as the correct control signal instead of $\tilde{u}_c[t]$.

4.2. Detector II for the Actuating Signal

In Figure 1, attacks ③ are targeted for actuating signals. The process for testing actuating signals includes the calculation of expected actuating signal $\bar{u}_a[t]$ and the test of the consistency between transmitted actuating signal $\tilde{u}_a[t]$ and expected actuating signal $\bar{u}_a[t]$. The following is the actuator model.

Actuator Model. Consider an actuator that can impose a physical influence on the physical process by executing the control signal. The expected actuating signal $\bar{u}_a[t]$ and the true actuating signal $u_a[t]$ at time t can be obtained by

$$\bar{u}_a[t] = h(\tilde{u}_c[t]), \tag{3}$$

$$u_a[t] = h(\tilde{u}_c[t]) + n[t] = \bar{u}_a[t] + n[t], \tag{4}$$

where $h(\cdot)$ denotes the modelling function for the actuator, $n[t] \in \mathbb{R}$ is the actuating noise. $\{n[t]\}$ has a zero mean independent identically distributed (i.i.d.) Gaussian sequence with 0 mean and $\sigma_n^2 \geq 0$ variance. Notably, true actuating signal $u_a[t]$ is only known to the attacker and unknown to the detector. For detector II, only transmitted actuating signal $\tilde{u}_a[t]$ can be obtained , which is possibly faulty, and expected actuating signal $\bar{u}_a[t]$ can be calculated to test the integrity of transmitted actuating signal $\tilde{u}_a[t]$.

Since the actuating signal is related to physical actions, unlike electrical signals, the difference between transmitted actuating signal $\tilde{u}_a[t]$ and expected actuating signal $\bar{u}_a[t]$ cannot be strictly zero. The computation of expected actuating signal $\bar{u}_a[t]$ and the verification of transmitted actuating signal $\tilde{u}_a[t]$ can be embedded in the downstream smart sensor. This integration aims to safeguard the integrity of actuating signals, particularly to prevent adverse effects on the physical process.

Define calculated actuating noise $\tilde{n}[t]$ as the difference between transmitted actuating signal $\tilde{u}_a[t]$ and expected actuating signal $\bar{u}_a[t]$, denoted by $\tilde{n}[t] := \tilde{u}_a[t] - \bar{u}_a[t]$. Therefore, we have the test for actuating signals.

Test for Actuating Signals. Check whether the calculated actuating noise sequence $\{\tilde{n}[t]\}$ belongs to a normal Gaussian distribution with 0 mean and σ_n^2 variance, denoted by $\{\tilde{n}[t]\} \sim \mathcal{N}(0, \sigma_n^2)$.

The test has an asymptotic form below:

I. $\{\tilde{n}[t]\}$ is Gaussian-distributed;

II. $\{\tilde{n}[t]\}$ is of 0 mean:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \tilde{n}[t] = 0; \tag{5}$$

III. $\{\tilde{n}[t]\}$ is of σ_n^2 variance:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \tilde{n}^2[t] = \sigma_n^2. \tag{6}$$

Here, T represents the total number of observational time steps or samples over which the system's behavior is analyzed. The condition where T approaches infinity, denoted by $\lim_{T\to\infty}$, assumes that our theoretical model conducts tests over an infinite time interval. However, real-world applications necessitate adjustments to accommodate practical limitations. Therefore, in practice, these tests are reduced to statistical tests over a finite time interval using sliding windows to continually analyze and refresh data segments.

Fault Tolerance. If transmitted actuating signal sequence $\{\tilde{u}_a[t]\}$ passes the corresponding statistical test, the flag indicating malicious actuating signal $flag_a[t]$ should be 0, which means the actuator is honest and reports the correct actuating signal, and vice versa. If $flag_a[t]$ is 1, the actuator should utilize expected actuating signal $\bar{u}_a[t]$, provided by the detector, to execute as the correct actuating signal, rather than relying on transmitted actuating signal $\tilde{u}_a[t]$. If the actuator is out of control, the system should be stopped in case of major safety accidents.

For a system without attacks, the transmitted actuating signal should be identically equal to the true actuating signal, denoted by $\tilde{u}_a[t] \equiv u_a[t]$; the calculated actuating noise should be identically equal to the true actuating noise, denoted by $\tilde{n}[t] \equiv n[t]$. Therefore, the distortion of actuating signal $\Delta_a[t] := \tilde{u}_a[t] - u_a[t]$, which is equal to the distortion of the actuating noise $\Delta_n[t] := \tilde{n}[t] - n[t]$, should be consistently 0 at every time step t. However, certain attacks might alter transmitted actuating signal sequence $\{\tilde{u}_a[t]\}$ within the noise range such that it bypasses detection. While such manipulations may lead to a non-zero noise distortion $\Delta_n[t]$, they are still considered ineffective since transmitted actuating signal sequence $\{\tilde{u}_a[t]\}$ maintains the same distribution as true actuating signal sequence $\{u_a[t]\}$.

For a system exposed to potential attacks, Theorem 1 ensures that any attack will be either detected or ineffective.

Theorem 1. *If the transmitted, possibly faulty, actuating signal sequence $\{\tilde{u}_a[t]\}$ passes test I. II. and III. in order to remain undetected, then:*

i. there are three possible values for the power of the distortion of the actuating signal sequence $\{\Delta_a[t]\}$:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \Delta_a^2[t] \in \{0, 2\sigma_n^2, 4\sigma_n^2\};$$

ii. the mean-square performance of true actuating signal sequence $\{u_a[t]\}$ is the same as the reported mean-square performance of transmitted actuating signal sequence $\{\tilde{u}_a[t]\}$:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} u_a^2[t] = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \tilde{u}_a^2[t]. \tag{7}$$

Proof. If transmitted actuating signal sequence $\{\tilde{u}_a[t]\}$ passes the test for actuating signals, the calculated actuating noise sequence $\{\tilde{n}[t]\}$ can be seen as another Gaussian noise sequence with 0 mean and σ_n^2 variance and

$$\tilde{u}_a[t] = h(\tilde{u}_c[t]) + \tilde{n}[t].$$

It is obvious that there are three possible values for the power of the distortion of the actuating signal sequence $\{\Delta_a[t]\}$:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \Delta_a^2[t] = \begin{cases} 0 & \{\tilde{n}[t]\} = \{n[t]\}. \\ 2\sigma_n^2 & \{\tilde{n}[t]\} \text{ and } \{n[t]\} \text{ are independent.} \\ 4\sigma_n^2 & \{\tilde{n}[t]\} = -\{n[t]\}. \end{cases}$$

The mean-square performance of true actuating signal sequence $\{u_a[t]\}$ is

$$\begin{aligned}\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}u_a^2[t] &= \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}(\bar{u}_a^2[t]+n^2[t]+2\bar{u}_a[t]n[t])\\ &= \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}(\bar{u}_a^2[t]+n^2[t])\\ &= \sigma_n^2+\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\bar{u}_a^2[t],\end{aligned}$$

since expected actuating signal $\bar{u}_a[t]$ and actuating noise $n[t]$ are independent. Similarly, the mean-square performance of transmitted actuating signal sequence $\{\tilde{u}_a[t]\}$ is

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\tilde{u}_a^2[t] = \sigma_n^2+\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\bar{u}_a^2[t],$$

since expected actuating signal $\bar{u}_a[t]$ and calculated actuating noise $\tilde{n}[t]$ are independent.

The proof is completed.

□

4.3. Detector III for the Sensor Reading

In Figure 1, attacks ④-⑥ can lead to malicious sensor readings. The process of testing sensor readings differs from the above two. Firstly, the transmitted actuating signal $\tilde{u}_a[t]$ and the sensor reading $\tilde{y}[t]$, serving as the system input and output respectively, are used to derive the Kalman filter's innovation $\tilde{z}[t]$ at time t. Then, the statistical properties of calculated innovation sequence $\{\tilde{z}[t]\}$ are tested to ensure the honesty of the sensor readings, thereby preventing the generation of erroneous control signals due to compromised sensor data.

Process Model. Consider a scalar linear time-invariant process described by

$$x[t+1] = ax[t]+b\tilde{u}_a[t]+w[t], \tag{8}$$

$$y[t] = cx[t]+d\tilde{u}_a[t]+v[t], \tag{9}$$

where $a,b,c,d\in\mathbb{R}$ are coefficients for the SISO control system; a is the state coefficient, affecting the system's next state; b is the input coefficient, determining how the input signal affects the state; c is the output coefficient, determining how the state variable influences the output; d is the direct transmission coefficient, indicating the direct impact of the input on the output without going through the system's dynamics. $x[t]\in\mathbb{R}$ denotes the system state at time t. $\tilde{u}_a[t]$ and $y[t]\in\mathbb{R}$ are the system input and output at time t, respectively. $w[t],v[t]\in\mathbb{R}$ are the process noise and measurement noise. $w[t]$ and $v[t]$ are zero-mean i.i.d. Gaussian noises with variances $\sigma_w^2\geq 0$ and $\sigma_v^2\geq 0$ respectively, independent of the initial state $x[0]$ of the system. $y[t]$ is the true measurement of the sensor at time t, which is known to the attacker and unknown to the detector. $\tilde{y}[t]$ is the transmitted measurement in the system at time t, which might be altered by a potential attacker. According to the system model, the true estimation of the system states can be obtained using true measurement $y[t]$ by the Kalman filter:

$$\begin{aligned} \hat{x}^-[t] &= a\hat{x}[t-1] + b\tilde{u}_a[t-1], && (10) \\ z[t] &= y[t] - c\hat{x}^-[t] - d\tilde{u}_a[t], && (11) \\ \hat{x}[t] &= \hat{x}^-[t] + kz[t], && (12) \end{aligned}$$

where $\hat{x}^-[t]$ and $\hat{x}[t]$ denote the true prior and posterior estimates of the system state $x[t]$; {z[t]} denotes the true innovation sequence, which is an i.i.d. Gaussian sequence with zero mean and variance $\sigma_z^2 = Pc^2 + \sigma_v^2$; $k = Pc/\sigma_z^2$ denotes the steady-state Kalman gain; P is the unique solution of the Riccati equation:

$$P = a^2P - a^2P^2c^2/(Pc^2 + \sigma_v^2) + \sigma_w^2.$$

However, the detector can only get transmitted measurement $\tilde{y}[t]$ instead of true measurement $y[t]$ to obtain the estimation of the system states by the Kalman filter:

$$\begin{aligned} \hat{\tilde{x}}^-[t] &= a\hat{\tilde{x}}[t-1] + b\tilde{u}_a[t-1], && (13) \\ \tilde{z}[t] &= \tilde{y}[t] - c\hat{\tilde{x}}^-[t] - d\tilde{u}_a[t], && (14) \\ \hat{\tilde{x}}[t] &= \hat{\tilde{x}}^-[t] + k\tilde{z}[t], && (15) \end{aligned}$$

where $\hat{\tilde{x}}^-[t]$ and $\hat{\tilde{x}}[t]$ denote the calculated prior and posterior estimates of the system state $x[t]$; $\{\tilde{z}[t]\}$ denotes the calculated innovation sequence. It is crucial to note that these calculated values are possibly faulty, underscoring the necessity of conducting tests on sensor readings to validate their integrity.

The statistical test for the calculated innovation sequence $\{\tilde{z}[t]\}$ can be deployed in the downstream controller to verify the integrity of sensor readings. The Kalman filter can be deployed within the controller as well or on the remote end.

Test for Sensor Readings. Check whether the calculated innovation sequence $\{\tilde{z}[t]\}$, which is possibly faulty, belongs to a normal Gaussian distribution with 0 mean and σ_z^2 variance, denoted by $\{\tilde{z}[t]\} \sim \mathcal{N}(0, \sigma_z^2)$. The test has an asymptotic form below:

I. $\{\tilde{z}[t]\}$ is Gaussian-distributed;

II. $\{\tilde{z}[t]\}$ is of 0 mean:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \tilde{z}[t] = 0; \tag{16}$$

III. $\{\tilde{z}[t]\}$ is of σ_z^2 variance:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \tilde{z}^2[t] = \sigma_z^2. \tag{17}$$

In practice, the above tests over an infinite time interval should be reduced to a finite time interval using sliding windows.

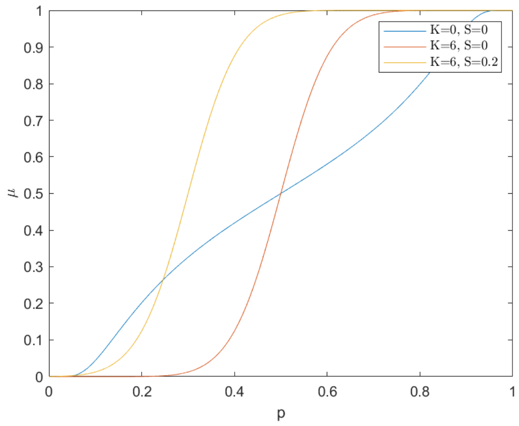


Figure 3: The μ-p function with different values of K and S. K and S are the scale and shift parameters, respectively.

Fault Tolerance. If calculated innovation sequence $\{\tilde{z}[t]\}$ passes the corresponding statistical test, the flag indicating malicious measurement $flag_s[t]$ should be 0, which means the sensor is honest and reports the correct sensor readings, and vice versa. If $flag_s[t]$ is 1, the detector will send expected measurement $\bar{y}[t] = c\hat{\tilde{x}}^-[t] + d\tilde{u}_a[t]$ instead of transmitted measurement $\tilde{y}[t]$ to the controller to generate correct control signals, which means we depend on the system model instead of the possibly faulty sensor readings to get the output. Given that the estimation of the system state is a recursive process, we attempt to mitigate the influence of possibly faulty sensor readings through the following modification of the Kalman filter:

$$\hat{\tilde{x}}[t] = \hat{\tilde{x}}^-[t] + \mu[t]k\tilde{z}[t], \tag{18}$$

where $\mu[t] \in [0, 1]$ is a self-adaptive coefficient that determines the extent to which the Kalman filter relies on possibly faulty sensor readings. A smaller value of $\mu[t]$ represents reduced reliance on these readings and enhanced fault tolerance of the system. However, this can lead to an increased false alarm rate (FAR) due to potential distortions in the original distribution of calculated innovation sequence $\{\tilde{z}[t]\}$. It is imperative to balance fault tolerance with FAR.

Hence, the value of self-adaptive coefficient $\mu[t]$ should be closer to 1 when sensor readings are more reliable and closer to 0 when sensor readings are more suspicious. The function determining self-adaptive coefficient $\mu[t]$ is given by:

$$\mu[t] = \begin{cases} \frac{1}{1+e^{-\lambda[t]}}, \lambda[t] = K \cdot tan(\pi(0.5 + p[t] + S)) & p[t] \in [0, 1-S), \\ 1 & p[t] \in [1-S, 1], \end{cases}$$

where $p[t] \in [0, 1]$ represents the p-value of the statistical test used in the detector III, which is positively correlated with the signal's reliability; K and S are the scale and shift hyperparameters to adjust the shape and position of the function for the desired self-adaptive effect of self-adaptive coefficient $\mu[t]$. Figure 3 shows the μ-p function with different values of K and S. By leveraging the $tan(\cdot)$ function, the domain of definition is mapped from $[0, 1]$ to $(-\infty, +\infty)$. Subsequently, the sigmoid function is employed to produce the S-shaped curve. This S-shape function is capable of distinguishing between two states with a gentle transition, which aligns with our objectives. The parameters K and S adjust the shape of the function and the 'boundary' between 'trust the sensor reading' and 'not trust the sensor reading', respectively.

For a system without attacks, the transmitted measurement should be identically equal to the true measurement, denoted by $\tilde{u}_a[t] \equiv u_a[t]$; the calculated innovation should be identically equal to the true innovation, denoted by $\tilde{z}[t] \equiv z[t]$; Therefore, the distortion of innovation $\Delta_z[t] := \tilde{z}[t] - z[t]$ should be consistently 0 at every time step t. Certain attacks altering calculated posterior estimates of the system state $\tilde{\hat{x}}[t]$ within the noise range to avoid detection are still considered ineffective as they have no impact on the control system performance.

For a system exposed to potential attacks, Theorem 2 ensures that any attack will be either detected or ineffective.

Theorem 2. *If the calculated, possibly faulty, innovation sequence $\{\tilde{z}[t]\}$ passes test I. II. and III. in order to remain undetected, then:*

i. there are three possible values for the power of the distortion of the innovation sequence $\{\Delta_z[t]\}$:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \Delta_z^2[t] \in \{0, 2\sigma_z^2, 4\sigma_z^2\};$$

ii. the mean-square performance of true posterior estimates of the system state sequence $\{\hat{x}[t]\}$ is the same as the reported mean-square performance of calculated posterior estimates of the system state sequence $\{\tilde{\hat{x}}[t]\}$:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \hat{x}^2[t] = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \tilde{\hat{x}}^2[t]. \tag{19}$$

Proof. If calculated innovation sequence $\{\tilde{z}[t]\}$ passes the test for sensor readings, it can be seen as a Gaussian noise sequence with 0 mean and σ_z^2 variance It is obvious that there are three possible values for the power of the distortion of the innovation sequence $\{\Delta_z[t]\}$:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \Delta_z^2[t] = \begin{cases} 0 & \{\tilde{z}[t]\} = \{z[t]\}, \\ 2\sigma_z^2 & \{\tilde{z}[t]\} \text{ and } \{z[t]\} \text{ are independent,} \\ 4\sigma_z^2 & \{\tilde{z}[t]\} = -\{z[t]\}. \end{cases}$$

Since calculated innovation equals true innovation plus the distortion, denoted by $\tilde{z}[t] = z[t] + \Delta_z[t]$, and from test III. for the measurement, shown in (17), we have

$$\begin{aligned} \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} (z[t] + \Delta_z[t])^2 &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} (z^2[t] + \Delta_z^2[t] + 2\Delta_z[t]z[t]) \\ &= \sigma_z^2. \end{aligned}$$

Since the variance of true innovation sequence $\{z[t]\}$ is σ_z^2, denoted by $\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} z^2[t] = \sigma_z^2$, we have

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} (\Delta_z^2[t] + 2\Delta_z[t]z[t]) = 0. \tag{20}$$

Define the distortion of the system state as the difference between the calculated and true posterior estimates of the system state, denoted by $\Delta_x[t] := \tilde{\hat{x}}[t] - \hat{x}[t]$. Note that

$$\begin{aligned} \Delta_x[t] &= a\tilde{\hat{x}}[t-1] + b\tilde{u}_a[t-1] + k\tilde{z}[t] - (a\hat{x}[t-1] + b\tilde{u}_a[t-1] + kz[t]) \\ &= a\Delta_x[t-1] + k\Delta_z[t]. \end{aligned}$$

Therefore, we have

$$\Delta_x[t] = \sum_{i=0}^{t-1} ka^i \Delta_z[t-i]. \tag{21}$$

Since the calculated posterior estimate of the system state equals the true posterior estimate of the system state plus the distortion, denoted by $\hat{\hat{x}}[t] = \hat{x}[t] + \Delta_x[t]$, we have

$$\begin{aligned}\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \hat{\hat{x}}^2[t] &= \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} (\hat{x}[t] + \Delta_x[t])^2 \\ &= \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} (\hat{x}^2[t] + \Delta_x^2[t] + 2\Delta_x[t]\hat{x}[t]).\end{aligned}$$

The mean of both true and calculated innovation sequences, denoted by $\{\tilde{z}[t]\}$ and $\{z[t]\}$, is 0. Therefore, we have

$$E\{\Delta_z[t]\} = E\{\tilde{z}[t]\} - E\{z[t]\} = 0.$$

Invoking the fact that the distortion of innovation sequence $\{\Delta_z[t]\}$ has a zero mean and independent of transmitted actuating signal sequence $\{\tilde{u}_a[t]\}$ and true posterior estimates of the system state sequence $\{\hat{x}[t]\}$, and substituting the system dynamics $\hat{x}[t] = a\hat{x}[t-1] + b\tilde{u}_a[t-1] + kz[t]$ and (21) into the following, we have

$$\alpha := \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} (\Delta_x^2[t] + 2\Delta_x[t]\hat{x}[t]).$$

$$\begin{aligned}\alpha &= \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} ((\sum_{i=0}^{t-1} ka^i\Delta_z[t-i])^2 + 2(\sum_{i=0}^{t-1} ka^i\Delta_z[t-i])(a\hat{x}[t-1] + b\tilde{u}_a[t-1] + kz[t])) \\ &= \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} ((\sum_{i=0}^{t-1} ka^i\Delta_z[t-i])^2 + 2(\sum_{i=0}^{t-1} ka^i\Delta_z[t-i])(\sum_{i=0}^{t-1} ka^i z[t-i])). \end{aligned} \tag{22}$$

We can expand the above two items,

$$(\sum_{i=0}^{t-1} ka^i\Delta_z[t-i])^2 = \sum_{i=0}^{t-1} k^2a^{2i}\Delta_z^2[t-i] + \sum_{i=0,j=0,i<j}^{t-1} 2k^2a^{(i+j)}\Delta_z[t-i]\Delta_z[t-j]; \tag{23}$$

$$(\sum_{i=0}^{t-1} ka^i\Delta_z[t-i])(\sum_{i=0}^{t-1} ka^i z[t-i]) = \sum_{i=0}^{t-1} k^2a^{2i}\Delta_z[t-i]z[t-i] + \sum_{i=0,j=0,i\neq j}^{t-1} k^2a^{(i+j)}\Delta_z[t-i]z[t-j]. \tag{24}$$

Given that the distortion of the innovation sequence $\{\Delta_z[t]\}$ has a zero mean, and each distortion of the innovation $\Delta_z[t]$ is independent of both the true innovation $z[k]$ and any other distortion in the innovation sequence $\Delta_z[k]$ for $k \neq t$, it follows that

$$\lim_{T\to\infty} \frac{1}{T}\sum_{i\neq j} \Delta_z[i]\Delta_z[j] = 0; \tag{25}$$

$$\lim_{T\to\infty} \frac{1}{T}\sum_{i\neq j} \Delta_z[i]z[j] = 0. \tag{26}$$

Table 4: Experiment setting.

Parameter/Component	Value/Model
Sampling frequency	100Hz
Input signal	a 1-Hz sinusoidal signal with 50 amplitude and 54° phase
Controller	a PI controller with proportional coefficient of 0.25 and integral coefficient of 0.48
Actuator	a 20-amplitude limiting module with a time step delay and actuating noise: $\sigma_n = 0.03$
Physical process	$a = 0.8, b = 1, c = 1.2, d = 0.3$, process noise: $\sigma_w = 0.08$, measurement noise: $\sigma_v = 0.02$
Window size	$\tau_a = \tau_s = 50$
Acceptable FAR	5% (for the χ^2 test)
Self-adaptive μ	scale parameter $K = 8.57$, shift parameter $S = 0.25$

Substituting (23), (24), (25), (26) and (20) into (22), we have

$$\begin{aligned}\alpha &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{i=0}^{t-1} k^2 a^{2i} (\Delta_z^2[t-i] + 2\Delta_z[t-i] z[t-i]) \\ &= 0.\end{aligned}$$

Therefore, we have

$$\begin{aligned}\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \hat{\tilde{x}}^2[t] &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} (\hat{x}[t] + \Delta_x[t])^2 \\ &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \hat{x}^2[t] + \alpha \\ &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \hat{x}^2[t].\end{aligned}$$

The proof is completed.

□

5. Experiments

This section demonstrates the experimental setup and findings. Detailed descriptions of the experiment settings are provided in subsection 5.1. The superior performance of isomorphic analysis compared to baseline methods is established in subsection 5.2. The case study section, found in subsection 5.3, illustrates the operational mechanics of our method and its effectiveness in maintaining stable system states under attack scenarios, thereby ensuring continued normal operation to a significant extent.

5.1. Setting

Table 4 details the technical specifications and parameters utilized in our experimental setup. We adopt a standard Kalman filter to estimate the system state and obtain the innovation sequence. In our work, we utilize the χ^2 test as the statistical method for detector II and detector III. The window sizes of the χ^2 test for actuating signals and sensor readings are τ_a and τ_s, respectively.

Table 5: Comparative experiment result. In the ‘Attack Type’ column, ‘Add’ and ‘Multi’ denote additive and multiplicative attacks, respectively. The evaluation metrics include detection rate (DR), root mean square error (RMSE), time delay (TD), and false alarm rate (FAR).

Method	Attack Type	Control Signals				Actuating Signals				Sensor Readings			
		DR	RMSE	TD	FAR	DR	RMSE	TD	FAR	DR	RMSE	TD	FAR
Isomorphic	Add	100%	0	0s	0	98%	5.54	0.038s	4.60%	99%	2.12	0.114s	5.09%
	Multi	100%	0	0s		99%	1.46	0.088s		99%	2.20	0.118s	
Watermark	Add	0	3.99	-	-	0	14.53	-	-	83%	5.90	0.114s	9.09%
	Multi	0	2.05	-		0	2.05	-		96%	3.24	0.151s	
OCSVM	Add	98%	5.29	0.1506s	0	96%	5.84	0.123s	0.17%	82%	7.61	0.208s	1.16%
	Multi	100%	2.66	0.179		100%	2.66	0.205		100%	4.18	0.097s	
IF	Add	93%	5.29	0.020s	0	93%	5.84	0.020s	0.33%	78%	7.61	0.061s	2.33%
	Multi	100%	2.66	0.294		97%	2.66	0.168		100%	4.18	0.108s	

5.2. Comparative Experiment

We compare the detection performance of our isomorphic analysis method not only with the representative white-box detection method, dynamic watermarking, but also with black-box detection methods, namely One-Class SVM (OCSVM) and IF. These compared methods are selected for their ability to be trained exclusively on normal data and subsequently utilized to differentiate between normal and anomalous data, aligning with the operational context of our approach. The comparison results, as presented in Table 5, highlight the superior detection and fault tolerance capabilities of our approach.

5.2.1. Attack Implementation

There are two kinds of FDI attacks that result in the distortion of transmitted signals in the system: one is additive and the other is multiplicative [26]. We implement both additive and multiplicative attacks on control signals, actuating signals, and sensor readings to test the comprehensive protection provided by detection methods.

Additive Attacks. Additive attacks involve the attacker distorting the signal by adding a fixed value each time. We conduct different additive attacks 100 times to evaluate detection performance. The additive attacks can be represented by

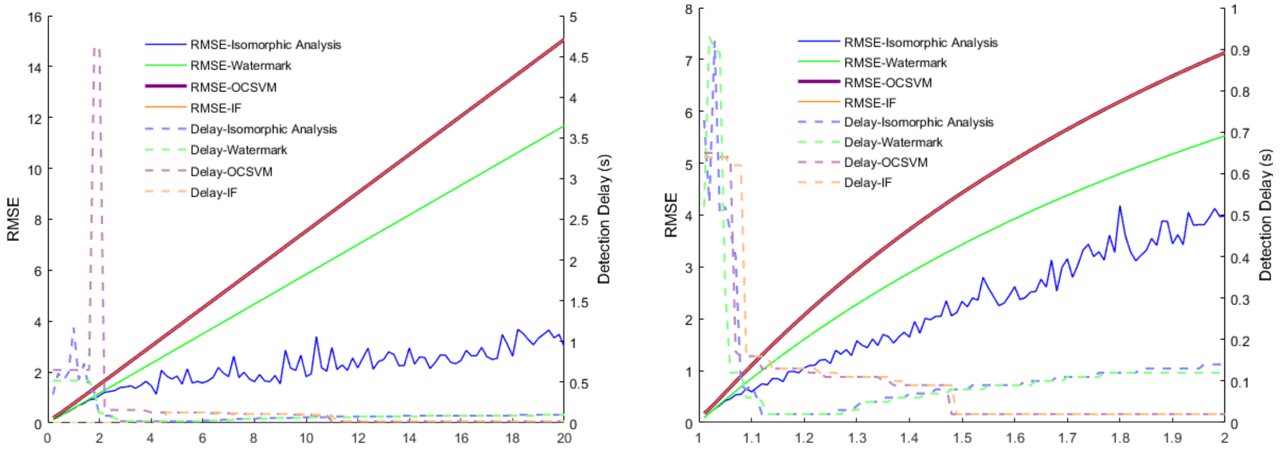
$$\widetilde{signal}[t] = signal[t] + add, \quad add \in (0 : 0.2 : 20],$$

where $signal[t]$ denotes the benign signal that has not been tampered with at time t, $\widetilde{signal}[t]$, $\widetilde{signal}[t] \in \{\tilde{u}_c[t], \tilde{u}_a[t], \tilde{y}[t]\}$, add denotes the value added to distort the signal, and $(0 : 0.2 : 20]$ indicates a sequence starting at 0 (not included), incrementing by 0.2, up to (and including) 20.

Multiplicative Attacks. Multiplicative attacks involve the attacker distorting the signal by multiplying it by a fixed value each time. We conduct different multiplicative attacks 100 times to evaluate the detection performance. The multiplicative attacks can be represented by

$$\widetilde{signal}[t] = signal[t] * multi, \quad multi \in (1 : 0.01 : 2],$$

where $multi$ denotes the multiplier for distorting the signal, and $(1 : 0.01 : 2]$ indicates a sequence starting at 1 (not included), incrementing by 0.01, up to (and including) 2.



(a) Additive attacks.

(b) Multiplicative attacks.

Figure 4: Comparative analysis of RMSE and detection time delay across attack amplitudes.

5.2.2. Evaluation Metrics and Performance Analysis

The comparative experiments evaluate the detection performance of control signals, actuating signals, and sensor readings under both additive and multiplicative attacks. The simulation time is 10s and all the attacks are launched at the start of 4s. Given that both our method and the watermarking technique are founded on statistical principles, we standardized the acceptable FAR threshold for the χ^2 test at 5% to ensure fairness.

False Alarm Rate (FAR). During the system's normal operation, the FAR serves as a critical metric, reflecting the reliability of the detection method. We assess the FAR by examining the likelihood of the detector falsely identifying an attack under normal operation of the system. Specifically, it is calculated as the percentage of total simulation samples in which the detector inaccurately signals an intrusion during our 10s simulation without any attack.

Our isomorphic analysis method employs three isomorphic detectors, enabling us to calculate separate FARs for control signals, actuating signals, and sensor readings. Similarly, the machine learning methods, OCSVM and IF, also employ separate detectors for each signal type. Conversely, the watermarking method utilizes only a single detector for sensor readings, yielding a corresponding FAR solely for this category, leading to its limited FAR representation.

For the detectors II and III utilizing the χ^2 statistical test, we set an acceptable FAR threshold of 5%. For these two detectors, the FARs in the experiment are close to 5%, underscoring the validity and theoretical consistency of our experimental outcomes. Notably, under additive attacks on sensor readings, our method presents a low FAR of 4.99%. Conversely, the watermarking method shows a higher FAR of 9.09% for the same scenario, indicating the superior robustness and the reduced susceptibility of our methodology to false alarms.

Machine learning methods like OCSVM and IF might demonstrate lower FARs due to their adaptive nature based on the training data. However, this adaptability may come at the expense of generalizability and could potentially overlook attack patterns that are not well-represented in the training set.

Detection Rate. In our experiments, we use the 'detection rate' to quantify the detection method's ability to identify attacks. Consider additive attacks as an example: the detection rate is calculated as the proportion of attacks successfully detected out of 100 additive attack implementations. For every implementation of the attack, we launch it at the start of 4s of the simulation time. Given that detectors can produce false alarms even during normal operation, we only recognize an attack as successfully detected if the count of alarms triggered by the detector during the attack exceeds the count of false alarms recorded over the same period when no attack is present. The same definition of detection rate applies to multiplicative attacks.

In assessing comparative performance, our isomorphic analysis method obviously excels with exceptional detection rates under both additive and multiplicative attacks on control signals, actuating signals, and sensor readings, consistently achieving over a 98% detection rate.

In contrast, the watermarking method fails to detect attacks on control and actuating signals altogether. Such shortcomings reveal design vulnerabilities, particularly an oversight of the cyber vulnerabilities present within each networked component of open ICSs.

Machine learning methods, OCSVM and IF, perform comparably to our isomorphic method for multiplicative attacks but show weaknesses for additive attacks, especially with sensor readings. This suggests that the subtle nature of changes introduced by additive attacks may present challenges for black-box models that are trained without an understanding of the control system's operational logic.

Therefore, our isomorphic analysis method provides comprehensive detection across all signal types and attack forms, underscoring its suitability for robust FDI attack detection in open ICSs.

Root Mean Square Error (RMSE). RMSE assesses the degree of distortion in the system state when subjected to an attack. For evaluation, we use the true system state during normal operation as the reference. The RMSE is then computed by comparing the state under attack with this normal state, providing a measure of the system's deviation from its expected state due to the attack.

As illustrated in Table 5, our method demonstrates superior performance by consistently presenting lower RMSE values compared to both the watermarking and machine learning methods for all types of attacks and signals. This superior mitigation of system state distortions is attributed to our method's post-attack corrective actions. Our approach proactively corrects malicious signals, thus enhancing the system's fault tolerance and resilience. This ability to correct and adapt in real time ensures that our system maintains operational normality, even under adverse conditions. In comparison, though sometimes effective, baseline methods lack this dynamic corrective capability, which is reflected in their generally higher RMSE values, especially under more sophisticated attack scenarios.

Time Delay. Time delay indicates the detector's responsiveness to attacks. Specifically, during the attack phase, it represents the interval between the onset of the attack and the detector's first alarm.

Our isomorphic analysis method exhibits a balanced performance, showing a slightly better time delay in multiplicative attacks when compared to the watermarking method, indicating a prompt detection capability that is crucial for immediate response and system recovery. Particularly for control and actuating signals, where the watermarking method fails to detect anomalies, our method ensures no gaps in security coverage.

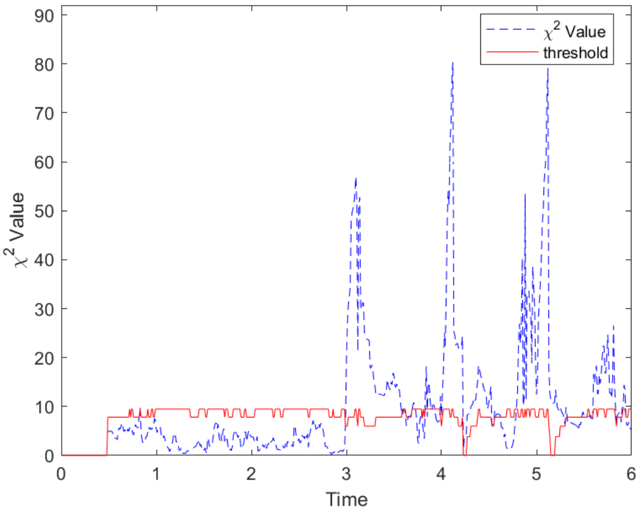


Figure 5: The threshold and χ^2 value under attack ($\mu = 0$).

While OCSVM and IF demonstrate quicker detection in certain scenarios, such as multiplicative attacks on sensor readings, their performance is not uniformly superior across all types of signals and attacks. This highlights an advantage of our isomorphic method: it provides a robust defense across various signal types without sacrificing much timeliness of detection.

Comparative Analysis of RMSE and Time Delay across Attack Amplitudes. To better demonstrate the performance of our isomorphic detection method relative to benchmark approaches, we present a detailed comparative analysis in Figure 4. This analysis depicts RMSE and detection time delay trends in response to varying amplitudes of both additive and multiplicative attacks.

As attack intensity increases, an expected rise in RMSE is observed across all methods, reflecting greater distortions in the system state caused by the more potent attacks. Conversely, detection time delay generally decreases with increasing attack strength, indicating faster detection as the attacks become less covert.

Our isomorphic analysis method not only achieves superior detection results but also excels in proactively correcting malicious signals, significantly mitigating system state distortions and minimizing the impact of attacks. While machine learning methods such as OCSVM and IF may exhibit rapid response in certain scenarios, they lack fault tolerance capabilities. Our method, in contrast, provides high fault tolerance and consistently maintains system integrity, even under adverse conditions.

The comparative experiment results clearly demonstrate our method's outstanding performance in FDI attack detection within ICSs, notably excelling in detection capabilities and fault tolerance. It outperforms both the dynamic watermarking method and machine learning benchmarks, ensuring comprehensive and reliable detection across a variety of attack types and providing robust system protection.

5.3. Case Study

In this section, we provide a case study of FDI attacks on sensor readings to demonstrate the effectiveness of our detection method. Additionally, we discuss the role of the parameter μ in resisting the adverse effects of attacks, namely the distortion of the system state.

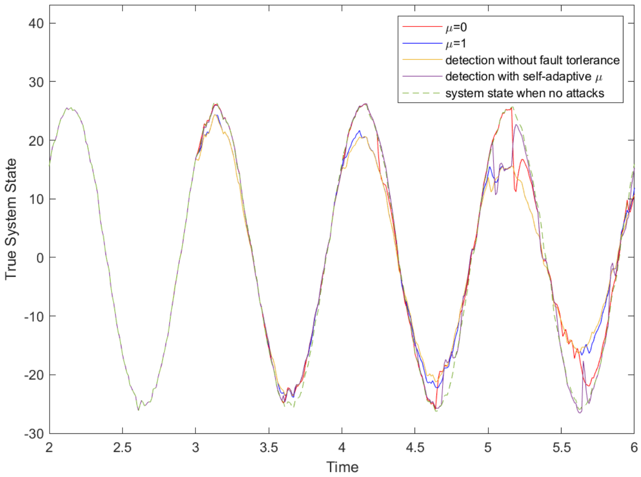


Figure 6: True system states under the attack and the normal condition.

Consider a scenario in which an attacker starts manipulation of the sensor readings at the moment 3s and gradually increases the intensity of the attack over time:

$$\tilde{y}[t] = \begin{cases} y & 0s \le t < 3s, \\ 1.15y & 3s \le t < 4s, \\ 1.4y & 4s \le t < 5s, \\ 2y & 5s \le t \le 6s. \end{cases}$$

Figure 5 plots χ^2 value and the corresponding threshold during the experiment. Obviously, the χ^2 value surpasses the threshold consistently after 3s, denoting the onset of an attack. Notably, peaks in χ^2 arise each time the attack pattern shifts, highlighting our method's sensitivity to the initial stage of the attack.

In (18), the coefficient μ is self-adaptive according to the condition of transmitted signals over time. Setting μ as a constant allows us to calibrate detectors to achieve different fault tolerance and FARs. In this context, (18) will be active only when the $flag_s$ is 1.

Figure 6 demonstrates the detector's resistance to distortion under different μ configurations, with our designed self-adaptive μ showing the best fault tolerance. Due to the recursive nature of the control system, the statistical detector cannot always achieve its theoretical performance, meaning it might not completely prevent distortions. Specifically, if the detector does not raise an alarm immediately after an attack, the system state can become distorted in the subsequent time step. Therefore, timely human intervention remains crucial upon attack detection.

6. Discussion

In this section, we first explore the minimal resource impact of our isomorphic detection method for supporting a cost-effective implementation in industrial settings. Following this, we address the scope and limitations of our detection capabilities regarding non-FDI attacks.

6.1. Resource Consumption

Enhancing a system's security usually necessitates increased resource consumption. The challenge lies in achieving an optimal balance between improved security and efficient resource utilization. Our isomorphic detectors exemplify this balance, merging enhanced protection with efficiency. All three proposed isomorphic detectors can be implemented as virtual models, utilizing existing industry resources and functioning on smart devices without the introduction of additional hardware. Upon deployment, these virtual models consume minimal computational costs due to their lightweight design and reliance on simple statistical tests. Additionally, the computation across these detectors can be parallelized to enhance efficiency. In contrast to the more resource-intensive, complex data-driven algorithms typical of machine learning or deep learning, this method offers significant cost savings.

For scenarios where controller failure could result in substantial risks or where high system availability is imperative, our system design allows for the use of redundant hardware. This adaptability enables the system configuration to be closely aligned with specific risk tolerances and resource capacities under different industrial environments.

Consequently, our isomorphic detectors provide an effective yet resource-conscious solution for enhancing the security of ICSs.

6.2. Limitations

Our methodology demonstrates robust effectiveness against FDI attacks, yet it is crucial to recognize that the cybersecurity threat landscape is vast, encompassing sophisticated tactics like Byzantine [27] and composite attacks [28]. Our detection design adeptly identifies signal alterations. However, it might not have the sensitivity required to detect activities that compromise privacy without altering data, such as passive eavesdropping. Additionally, scenarios where devices report signals that do not accurately represent their operational state, such as in replay attacks where previously captured legitimate signals are resent, pose a challenge to our detection methodology.

Moreover, it is important to note that the theoretical proof of our method's efficacy is based on an infinite time interval assumption. In practice, however, we are constrained to a finite sliding window for detection. This limitation creates a potential exploit for covert attacks [29], which could leverage the finite window to remain undetected by operating below our system's detection threshold.

7. Conclusion

In this paper, we introduce an innovative isomorphic analysis detection method for open ICSs, offering robust protection against FDI attacks. By fully utilizing system information, we construct three isomorphic detectors that together cover all critical components of the physical control system, addressing vulnerabilities that are often overlooked by methods focusing solely on sensor-based attacks. The proposed three isomorphic detectors not only identify FDI attacks but also undertake corrective measures to ensure the transmission of more dependable signals to downstream components, thereby mitigating the impact of attacks. The effectiveness of our method is supported by theoretical validation and empirical evidence. Through experimental evaluations, our technique outperforms baselines such as the dynamic watermarking method, SVM, and IF models across crucial metrics, including detection accuracy, fault tolerance, FAR, and detection time delay. While our method excels in the context of FDI attacks, we acknowledge its limitations in detecting more sophisticated cyber threats that may not involve direct signal manipulation. In the future, we plan to extend our isomorphic analysis detection method to multiple output-multiple input (MIMO) and nonlinear systems for wider applicability.

Acknowledgements

This research is supported in part by the National Natural Science Foundation of China (Grant No. 92067109, 61873119, 62211530106), in part by Shenzhen Science and Technology Program (Grant No. ZDSYS20210623092007023, GJHZ20210705141808024), in part by the Educational Commission of Guangdong Province (Grant No. 2019KZDZX1018), and in part by the UGC General Research Fund (Grant No. 17203320, 17209822) from Hong Kong..

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used ChatGPT to improve the readability and language of the manuscript. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

[1] I. C. S. C. E. R. Team, Recommended practice: improving industrial control systems cyber security with defense-in-depth strategies, Tech. rep., Homeland Security (September 2016).

[2] A. A. Cardenas, S. Amin, S. Sastry, Secure control: Towards survivable cyber-physical systems, in: 2008 The 28th International Conference on Distributed Computing Systems Workshops, 2008, pp. 495–500.

[3] K. J. Houle, G. M. Weaver, Trends in denial of service attack technology, cert and cert coordination center (2001).

[4] A. Householder, A. Manion, L. Pesante, G. M. Weaver, R. Thomas, Managing the threat of denial-of-service attacks, CERT Coordination Center 10 (2001).

[5] J. Ye, L. Guo, B. Yang, F. Li, L. Du, L. Guan, W. Song, Cyber–physical security of powertrain systems in modern electric vehicles: Vulnerabilities, challenges, and future visions, IEEE Journal of Emerging and Selected Topics in Power Electronics 9 (4) (2021) 4639–4657.

[6] Recent advances on filtering and control for cyber-physical systems under security and resource constraints, Journal of the Franklin Institute 353 (11) (2016) 2451–2466.

[7] M. Esmalifalak, L. Liu, N. Nguyen, R. Zheng, Z. Han, Detecting stealthy false data injection using machine learning in smart grid, IEEE Systems Journal 11 (3) (2017) 1644–1652.

[8] S. A. Foroutan, F. R. Salmasi, Detection of false data injection attacks against state estimation in smart grids based on a mixture gaussian distribution learning method, IET Cyber-Physical Systems: Theory & Applications 2 (4) (2017) 161–171.

[9] C. Feng, T. Li, D. Chana, Multi-level anomaly detection in industrial control systems via package signatures and lstm networks, in: 2017 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), 2017, pp. 261–272.

[10] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, C. Eickhoff, A transformer-based framework for multivariate time series representation learning, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery &amp; Data Mining, KDD '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 2114–2124.

[11] B. Satchidanandan, P. R. Kumar, Dynamic watermarking: Active defense of networked cyber–physical systems, Proceedings of the IEEE 105 (2) (2017) 219–240.

[12] C. Fung, S. Srinarasi, K. Lucas, H. B. Phee, L. Bauer, Perspectives from a comprehensive evaluation of reconstruction-based anomaly detection in industrial control systems, in: V. Atluri, R. Di Pietro, C. D. Jensen, W. Meng (Eds.), Computer Security – ESORICS 2022, Springer Nature Switzerland, Cham, 2022, pp. 493–513.

[13] Z. Ding, M. Fei, An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window, IFAC Proceedings Volumes 46 (20) (2013) 12–17, 3rd IFAC Conference on Intelligent Control and Automation Science ICONS 2013.

[14] C. Feng, V. R. Palleti, A. Mathur, D. Chana, A systematic framework to generate invariants for anomaly detection in industrial control systems., 2019.

[15] A. Deng, B. Hooi, Graph neural network-based anomaly detection in multivariate time series, Proceedings of the AAAI Conference on Artificial Intelligence 35 (5) (2021) 4027–4035.

[16] A. Sargolzaei, K. Yazdani, A. Abbaspour, C. D. Crane III, W. E. Dixon, Detection and mitigation of false data injection attacks in networked control systems, IEEE Transactions on Industrial Informatics 16 (6) (2020) 4281–4292.

[17] Q. Su, S. Li, Y. Gao, X. Huang, J. Li, Observer-based detection and reconstruction of dynamic load altering attack in smart grid, Journal of the Franklin Institute 358 (7) (2021) 4013–4027.

[18] Z. Abdollahi Biron, S. Dey, P. Pisu, Real-time detection and estimation of denial of service attack in connected vehicle systems, IEEE Transactions on Intelligent Transportation Systems 19 (12) (2018) 3893–3902.

[19] W. Ao, Y. Song, C. Wen, Adaptive cyber-physical system attack detection and reconstruction with application to power systems, IET Control Theory & Applications 10 (12) (2016) 1458–1468.

[20] K. Manandhar, X. Cao, F. Hu, Y. Liu, Detection of faults and attacks including false data injection attack in smart grid using kalman filter, IEEE Transactions on Control of Network Systems 1 (4) (2014) 370–379.

[21] Y. Chen, T. Li, Y. Long, W. Bai, Attacks detection and security control for cyber-physical systems under false data injection attacks, Journal of the Franklin Institute 360 (14) (2023) 10476–10498.

[22] C. Hu, P. Fan, Y. Li, I.-J. Chiu, Y. Wang, Y. Zhou, Y. Li, H. Li, False data injection attack detection of cyber-physical charging systems based on time-frequency analysis, in:

2023 International Conference on Smart Electrical Grid and Renewable Energy (SEGRE), 2023, pp. 80–88.

[23] M. Porter, P. Hespanhol, A. Aswani, M. Johnson-Roberson, R. Vasudevan, Detecting generalized replay attacks via time-varying dynamic watermarking, IEEE Transactions on Automatic Control 66 (8) (2021) 3502–3517.

[24] R. Deng, G. Xiao, R. Lu, Defending against false data injection attacks on power system state estimation, IEEE Transactions on Industrial Informatics 13 (1) (2017) 198–207.

[25] J. Hao, R. J. Piechocki, D. Kaleshi, W. H. Chin, Z. Fan, Sparse malicious false data injection attacks and defense mechanisms in smart grids, IEEE Transactions on Industrial Informatics 11 (5) (2015) 1–12.

[26] H. Zhu, L. Xu, Z. Bao, Y. Liu, L. Yin, W. Yao, C. Wu, L. Wu, Secure control against multiplicative and additive false data injection attacks, IEEE Transactions on Industrial Cyber-Physical Systems 1 (2023) 92–100.

[27] X. Gong, X. Li, Z. Shu, Z. Feng, Resilient output formation-tracking of heterogeneous multiagent systems against general byzantine attacks: A twin-layer approach, IEEE Transactions on Cybernetics (2023) 1–13.

[28] X. Gong, M. V. Basin, Z. Feng, T. Huang, Y. Cui, Resilient time-varying formation-tracking of multi-uav systems against composite attacks: A two-layered framework, IEEE/CAA Journal of Automatica Sinica 10 (4) (2023) 969–984.

[29] J. He, X. Gong, Resilient path planning of unmanned aerial vehicles against covert attacks on ultrawideband sensors, IEEE Transactions on Industrial Informatics 19 (11) (2023) 10892–10900.

Appendix A. An Alternative Proof Strategy for Theorem 1

An alternative proof approach for Theorem 1 exists, which primarily aims to validate (7). While this way is more complex, it serves as the basic proof idea for Theorem 2.

Since calculated actuating noise equals true actuating noise plus the distortion, denoted by $\tilde{n}[t] = n[t] + \Delta_n[t]$, and from test III. for the actuating signal, shown in (6), we have

$$\begin{aligned} \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} (n[t] + \Delta_n[t])^2 &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} (n^2[t] + \Delta_n^2[t] + 2\Delta_n[t]n[t]) \\ &= \sigma_n^2. \end{aligned}$$

Since the variance of actuating noise sequence $\{n[t]\}$ is σ_n^2, denoted by $\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} n^2[t] = \sigma_n^2$, we have

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} (\Delta_n^2[t] + 2\Delta_n[t]n[t]) = 0. \tag{A.1}$$

Since transmitted actuating signal is equal to true actuating signal plus the distortion, denoted by $\tilde{u}_a[t] = u_a[t] + \Delta_a[t] = u_a[t] + \Delta_n[t]$, we have

$$\begin{aligned}\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\tilde{u}_a^2[t] &= \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}(u_a[t]+\Delta_n[t])^2\\ &= \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}(u_a^2[t]+\Delta_n^2[t]+2\Delta_n[t]u_a[t]).\end{aligned}$$

The mean of both true and calculated actuating noise sequences, denoted by $\{\tilde{n}[t]\}$ and $\{n[t]\}$, is 0. Therefore, we have

$$E\{\Delta_n[t]\} = E\{\tilde{n}[t]\} - E\{n[t]\} = 0.$$

Hence, invoking the fact that the distortion of the actuating noise sequence $\{\Delta_n[t]\}$ is of 0 mean and independent of $\{h(\tilde{u}_c[t])\}$, we have

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\Delta_n[t]h(\tilde{u}_c[t]) = 0. \tag{A.2}$$

Noting that true actuating signal $u_a[t] = \bar{u}_a[t] + n[t]$ and (A.1), we have

$$\begin{aligned}\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}(\Delta_n^2[t]+2\Delta_n[t]u_a[t]) &= \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}(\Delta_n^2[t]+2\Delta_n[t](\bar{u}_a[t]+n[t]))\\ &= \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}(\Delta_n^2[t]+2\Delta_n[t]n[t])\\ &= 0.\end{aligned}$$

Therefore, we have

$$\begin{aligned}\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\tilde{u}_a^2[t] &= \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}(u_a[t]+\Delta_n[t])^2\\ &= \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}u_a^2[t] + \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}(\Delta_n^2[t]+2\Delta_n[t]u_a[t])\\ &= \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}u_a^2[t].\end{aligned}$$

The proof is completed.