A Game-Theoretic Approach to Secure Estimation and Control for Cyber-Physical Systems with a Digital Twin

1st Zhiheng Xu
School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore
Email: zhiheng.xu688@gmail.com

2nd Arvind Easwaran
School of Computer Science and Engineering
Nanyang Technological University, Singapore
Email: arvinde@ntu.edu.sg

Abstract—Cyber-Physical Systems (CPSs) play an increasingly significant role in many critical applications. These valuable applications attract various sophisticated attacks. This paper considers a stealthy estimation attack, which aims to modify the state estimation of the CPSs. The intelligent attackers can learn defense strategies and use clandestine attack strategies to avoid detection. To address the issue, we design a Chi-square detector in a Digital Twin (DT), which is an online digital model of the physical system. We use a Signaling Game with Evidence (SGE) to find the optimal attack and defense strategies. Our analytical results show that the proposed defense strategies can mitigate the impact of the attack on the physical estimation and guarantee the stability of the CPSs. Finally, we use an illustrative application to evaluate the performance of the proposed framework.

Index Terms—Cyber-Physical Systems, Data-Integrity Attack, Signaling Game with Evidence, Digital Twin.

I. INTRODUCTION

Cyber-Physical Systems (CPSs) integrate physical components (e.g., sensors, actuators, and controllers), computational resources, and networked communications [1]. The integration with networked communications highly enhances the flexibility and scalability of CPSs in various applications, such as large-scale manufacturing systems [2], intelligent transport systems [3], and smart grid infrastructure [4]. However, the valuable applications of CPSs attract many sophisticated attacks. A well-known example is the Stuxnet, a computer virus compromises Supervisory control and data acquisition (SCADA) of industrial systems [5]. Besides, we have witnessed many other control-system-related attacks, such as Maroochy Water attack [6], Unmanned Aerial Vehicle’s (UAV) GPS spoofing attack [7], and German Steel Mill cyber attack [8].

Due to the increasing number of security-related incidents in CPSs, many researchers have studied the features of these attacks and developed relevant defense strategies. Among many attack models, we focus on data-integrity attacks, where the attackers modify the original data used by the system or inject unauthorized data to the system [9]. The data-integrity attacks can cause catastrophic consequences in CPSs. For instance, the fake data may deviate the system to a dangerous trajectory or make the system oscillate with a significant amplitude, destabilizing the system. Therefore, how to mitigate the impact of data-integrity attacks becomes a critical issue in the security design of CPSs.

One typical data-integrity attack for CPSs is the Sensor-and-Estimation (SE) attack, where the attackers tamper the sensing or estimated information of the CPSs [10]. Given the SE attack, Fawzi et al. [11] have studied a SE attack and proposed algorithms to reconstruct the system state when the attackers corrupt less than half of the sensors or actuators. Miroslav Pajic et al. [12] have extended the attack-resilient state estimation for noisy dynamical systems. Based on Kalman filter, Chang et al. [13] have extended the secure estimation of CPSs to a scenario in which the set of attacked nodes can change over time. However, to recover the estimation, the above work requires a certain number of uncorrupted sensors or a sufficiently long time. Those approach might introduce a non-negligible computational overhead, which is not suitable for time-critical CPSs, e.g., real-time systems. Besides, since all the sensors do not have any protection, the attacker might easily compromise the large number of sensors, violating the assumptions of the above work.

Instead of recovering the estimation from SE attacks, researchers and operators also focus on attack detection [14]. However, detecting a SE attack could be challenging since the attackers’ strategies become increasingly sophisticated. Pasqualetti et al. [15] have identified the conditions of undetectable SE attacks for CPSs. Using the conventional statistic detection theory, e.g., Chi-square detection, may fail to discover an estimation attack if the attackers launch a stealthy attack [16]. Yuan Chen et al. [17] have developed optimal attack strategies, which can deviate the CPSs subject to detection constraints. Hence, the traditional detection theory may not sufficiently address the stealthy attacks in which the attackers can acquire the defense information. Besides, using classical cryptography to protect CPSs will introduce significant overhead, degrading the control performance of delay-sensitive CPSs [18].

The development of Digital Twin (DT) provides essential resources and environments for detecting sophisticated attacks. A DT could be a virtual prototype of a CPS, reflecting partial or entire physical processes [19]. Based on the concept of DT, researchers have developed numerous applications for CPSs [20]. The most relevant application to this paper is using a DT to achieve fault detection [21]. Similarly, we use the DT

This work was supported by Delta-NTU Corporate Lab for Cyber-Physical Systems with funding support from Delta Electronics Inc. and the National Research Foundation (NRF) Singapore under the Corp Lab@University Scheme.
to monitor the estimation process, mitigating the influence of the SE attack.

In this paper, we focus on a stealthy estimation attack, where the attackers know the defense strategies and aim to tamper the estimation results without being detected. Fig. 1 illustrates the underlying architecture of the proposed framework. To withstand the attack, we design a Chi-square detector, running in a DT. The DT connects to a group of protected sensing devices, collecting relevant evidence. We use cryptography (e.g., digital signature or message authentication code) to preserve the evidence from the attack. Hence, the DT can use the evidence to monitor the estimation of the physical systems. The cryptographic overhead will not affect physical performance since the execution of the real plant does not depend on the DT.

Different from the work [11]–[13], we have designed two independent channels, i.e., one is protected by standard cryptography, and the other one is the general sensing channel. Fig. 2 illustrates the structure of the framework. The main advantage of this structure is the cryptographical overhead will affect the control performance of the physical system due to the independency between these two channels.

![Diagram](image)

Fig. 2. The Architecture of the Framework: the Digital Twin has two channels to obtain the sensing information of the plant: one is secure, and the other is not secure. The secure channel provides less accurate data with a heavy computational overhead. DT only uses the secure channel to run the verification, and the computational overhead has negligible impact on the physical performance since these two channels are independent.

To analyze whether a stealthy attack can bypass DT’s detector, we use game theory to find the optimal attack and defense strategies. Game theory has been an essential tool in designing security algorithms since we can use it to search for the optimal defense strategies against intelligent attackers [22]. One related game model that can capture the detection issue is the Signaling Game (SG) [23]. However, instead of using the SG, we use a Signaling Game with Evidence (SGE), presented in [24], to protect the system from the attack. In an SGE, the DT’s detector will provide critical evidence to explore the stealthy attack. After integrating DT’s detector with CPSs, we use an SGE to study the stealthy attack and develop optimal defense strategies based on the equilibria of the SGE. Our analytical results show that the stealthy attackers have to sacrifice the impact on the physical system to avoid detection.

We organize the rest of the paper as follows. Section II presents the problem formulation in which we identify the control problem, attack model, and a signaling game framework. Section III analyzes the equilibrium of the game framework and the stability of the CPSs. Section IV uses the experimental results to evaluate the performance of the proposed defense mechanism. Finally, Section V concludes the entire paper.

II. SYSTEM MODELING AND CHARACTERIZATION

In this section, we first introduce the dynamic model of a CPS. Secondly, we define a stealthy estimation attack model. Based on a Digital Twin (DT), we design a Chi-square detector to monitor the estimation process. Finally, we define a Signaling Game with Evidence (SGE) to characterize the features of the stealthy attack.

A. System Model and Control Problem of a CPS

Suppose that the physical layer of a CPS is a control system. We assume that the control system can be linearized as a linear discrete-time system, given by

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

$$ y_k = Cx_k + v_k, $$

where $k \in \mathbb{Z}_+$ is the discrete-time instant; $x_k \in \mathbb{R}^{n_x}$ is the system state with an initial condition $x_0 \sim \mathcal{N}(0, \Sigma_x)$, and $\Sigma_x \in \mathbb{R}^{n_x \times n_x}$ is the covariance matrix; $u_k \in \mathbb{R}^{n_u}$ is the control input; $y_k \in \mathbb{R}^{n_y}$ is the sensor output; $w_k \in \mathbb{R}^{n_w}$ and $v_k \in \mathbb{R}^{n_v}$ are additive zero-mean Gaussian noises with covariance matrices $\Sigma_w$ and $\Sigma_v$ with proper dimensions; $A$, $B$, and $C$ are constant matrices with proper dimensions.

![Diagram](image)

Fig. 1. The Stealthy Estimation Attack and the Defense Mechanism based on a Digital Twin: the attacker aims to modify the estimation results in a CPS, while the Digital Twin protects the system by monitoring the results.

Given system model (1), we design a control policy $\mu : \mathbb{R}^{n_x} \to \mathbb{R}^{n_u}$ by minimizing the following expected linear quadratic cost function, i.e.,

$$ J_{LQG} = \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} \left( x_k^T Q x_k + u_k^T R u_k \right), $$

where $Q \in \mathbb{R}^{n_x \times n_x}$ and $R \in \mathbb{R}^{n_u \times n_u}$ are positive-definite matrices.

Note that the controller cannot use state $x_k$ directly, i.e., we need to design an observer to estimate $x_k$. Hence, minimizing function (3) is a Linear Regulator Gaussian (LQG) problem.
According to the separation principle, we can design the controller and state estimator separately. The optimal control policy \( \mu : \mathbb{R}^{n_z} \rightarrow \mathbb{R}^{n_u} \) is given by

\[
\mu_k(x_k) := K_k x_k,
\]
with \( K_k := -(R + B^T V_k B) B^T V_k \), where \( V_k \in \mathbb{R}^{n_x \times n_x} \) is the solution to the linear discrete-time algebraic Riccati equation

\[
V_{k+1} = Q + A^T V_k (A - BK_k), \quad \text{with} \quad V_0 = I_{n_x},
\]
and \( I_{n_x} \in \mathbb{R}^{n_x \times n_x} \) is an identity matrix.

We assume that \( (A, B) \) is stabilizable. Then, \( V_k \) will converge to a constant matrix \( V \) when \( k \) goes to infinity, i.e.,

\[
\lim_{k \to \infty} V_k = V, \quad \text{and} \quad \lim_{k \to \infty} \mu_k(x) = \mu(x) = K x.
\]

In the next subsection, we will use a Kalman filter to estimate \( x_k \) such that controller (4) uses this estimated value to control the physical system.

### B. Kalman Filter Problem

To use controller (4), we need to design an estimator. Let \( \hat{x}_k \in \mathbb{R}^{n_x} \) be the estimation of \( x_k \) and \( \hat{e}_k := \hat{x}_k - x_k \) be the error of the estimation. Given the observation \( y_{k-1} := \{y_0, y_1, \ldots, y_{k-1}\} \), we aim to solve the following Kalman filtering problem, i.e.,

\[
\min_{\hat{x}_k \in \mathbb{R}^{n_x \times n_x}} \mathbb{E}[|\hat{x}_k - x_k|^2 | y_{k-1}].
\]

To solve (6), we need the following lemma, which characterizes a conditioned Gaussian distribution.

**Lemma 1:** \([25]\) If \( a \in \mathbb{R}^{n_a}, b \in \mathbb{R}^{n_b} \) are jointly Gaussian with means \( \bar{a}, \bar{b} \) and covariances \( \Sigma_a, \Sigma_b, \) and \( \Sigma_{ab} = \Sigma_{ba}^T \), then given \( b \), distribution \( a \) is a Gaussian with

\[
\mathbb{E}[a | b] = \bar{a} + \Sigma_{ab} \Sigma_b^{-1} (b - \bar{b}),
\]

\[
\text{Cov}[a | b] = \Sigma_a - \Sigma_{ab} \Sigma_b^{-1} \Sigma_{ba}.
\]

To use Lemma 1 we define the covariance matrix of \( \hat{e}_k \) as

\[
\hat{P}_k := \mathbb{E}[\hat{e}_k \hat{e}_k^T | y_{k-1}] = \mathbb{E}[(\hat{x}_k - x_k)(\hat{x}_k - x_k)^T | y_{k-1}],
\]

with \( \hat{P}_0 = \Sigma_x \). Using the results of Lemma 1, we compute the optimal estimation iteratively, i.e.,

\[
\hat{x}_k = A_K \hat{x}_{k-1} + \hat{L}_k (y_{k-1} - C \hat{x}_{k-1}),
\]

where \( A_K := (A + BK) \). Gain matrix \( \hat{L}_k \in \mathbb{R}^{n_x \times n_y} \) and covariance \( \hat{P}_k \in \mathbb{R}^{n_x \times n_x} \) are updated using

\[
\hat{L}_k := A_K \hat{P}_k C^T (\Sigma_w + C \hat{P}_k C^T)^{-1},
\]

\[
\hat{P}_{k+1} := \Sigma_w + (A_K - \hat{L}_k C) \hat{P}_k A_K^T.
\]

Assuming that \( (A, C) \) is detectable, we can obtain that

\[
\lim_{k \to \infty} \hat{P}_k = \hat{P}^*,
\]

where \( \hat{P}^* \in \mathbb{R}^{n_x \times n_x} \) is a constant positive-definite matrix.

### C. Stealthy Estimation Attack

CPSs face an increasing threat in recent years. Numerous attack models for CPSs or networked control systems (NCSs) have been introduced in \([26]\). Among those attacks, one major attack is the data-integrity attack, where the attacker can modify or forge data used in the CPSs. For example, Liu et al. \([27]\) have studied a false-data-injection attack that can tamper the state estimation in electrical power grids.

![Fig. 3. The Stealthy Estimation Attack: intelligent attacker aims to deviate the state by modifying the estimation results.](image)

To mitigate the impact of a data-integrity attack on the state estimation, researchers have designed a Chi-square detector to monitor the estimation process. However, Yi Lin Mo et al. \([16]\) have analyzed stealthy integrity attack, which can pass the Chi-square detector by manipulating the perturbation of the injected data. Therefore, the main challenge is that the conventional fault detectors fail to protect the system from a stealthy attack. One real attack that can achieve the objective is the Advanced Persistent Threat (APT) \([28]\), which can compromise a cyber system by executing a zero-day exploration to discover the vulnerabilities.

In our work, we consider an intelligent attacker who can launch a stealthy estimation attack to tamper the estimation results. Fig. 3 illustrates how the attacker achieves its objective. The attacker can either modify the data in the sensors or the data in the estimator. Besides tampering the estimation results, the attacker is also aware of the intrusion detector. The attacker can know the defense strategy and play a stealthy attack to remaind unknown.

In the next subsection, based on the Digital Twin (DT), we design a cyber defender to withstand the stealthy estimation attack and discuss the benefits introduced by the DT. After presenting the game model, we will discuss the optimal defense strategies explicitly in Section III.

### D. Digital Twin for the CPS

As mentioned above, an intelligent attacker can learn the defense strategy and launch a stealthy estimation attack, which can modify the estimation results without being detected by the conventional detector, e.g., a Chi-square detector. To resolve the issue, we aim to design an advanced detector based on a Digital Twin (DT). After that, we use a game-theoretical approach to develop an optimal defense strategy.
Given the system information, we design a DT with the following dynamics:

\[
\hat{x}_k = A_k \hat{x}_{k-1} + \tilde{L}_k (z_{k-1} - D \hat{x}_{k-1}),
\]

\[
z_k = D \hat{x}_k + d_k,
\]

where \(\hat{x}_k \in \mathbb{R}^{n_x}\) is the DT’s estimation of state \(x_k\); \(z_k \in \mathbb{R}^{n_z}\) is the DT’s observation; \(D \in \mathbb{R}^{n_z \times n_x}\) is a constant matrix, \(d_k\) is a Gaussian noise with a covariance matrix \(\Sigma_d \in \mathbb{R}^{n_z}\).

Similar to problem [6], we compute \(\tilde{L}_k\) using the following iterations:

\[
\tilde{L}_k = A_k \tilde{P}_k D^T (\Sigma_d + D \tilde{P}_k D^T)^{-1},
\]

\[
\tilde{P}_{k+1} = \Sigma_w + (A_k - \tilde{L}_k D) \tilde{P}_k A_k^T,
\]

where \(\tilde{P}_0 = \Sigma_e\) and \(\tilde{P}_k\) is defined by

\[
\tilde{P}_k := \mathbb{E}[(\hat{x}_k - x_k)(\hat{x}_k - x_k)^T | z_{k-1}] = \mathbb{E}[\hat{e}_k \hat{e}_k^T | z_{k-1}],
\]

where \(\hat{e}_k := \hat{x}_k - x_k\) is the DT’s estimation error. We also assume that \((A_k, D)\) is detectable, i.e., we have

\[
\lim_{k \to \infty} \tilde{P}_k = \tilde{P}^*.
\]

Fig. 4 illustrates the architecture of a CPS with a DT. We summarize the main differences between Kalman filter [7] and the DT’s estimator [8] as follows. Firstly, the Kalman filter will use all available sensing information \(y_k\) to obtain estimation \(\hat{x}_k\). While the DT’s estimator just uses a minimum sensing information \(z_k \in \mathbb{R}^{n_z}\) to predict \(x_k\) as long as \((A, D)\) is detectable, i.e., \(n_y \geq n_x\). This feature reduces the dimension of \(z_k\), making it easier to protect \(z_k\). Secondly, we do not require a high accuracy for \(z_k\), since we only use \(z_k\) for attack detection. Hence, in general, \(\tilde{P}^*\) and \(\tilde{P}^*\) satisfy the condition that \(\text{tr} (\tilde{P}^*) \leq \text{tr} (\tilde{P}^*)\), where \(\text{tr} (P)\) is the trace of matrix \(P\).

Thirdly, we do not use any cryptography to protect \(y_k\) since the overhead introduced by the encryption scheme will degrade the performance of the physical system. However, we can use cryptography, such as Message Authentication Code (MAC) [29] or Digital Signature (DS) [30], to protect the integrity of \(z_k\). The overhead caused by the cryptography will not affect the physical system, because it does depend on \(z_k\). Besides, we can put the DT into a supercomputer or a cloud to resolve the overhead issue.

To sum up, \(z_k\) is an observation that is less accurate but more secure than \(y_k\). Given the distinct features of \(y_k\) and \(z_k\), we use \(y_k\) to estimate \(x_k\) for the physical control and use \(z_k\) for the detection in the DT.

Given DT’s estimator, we construct a Chi-square detector to monitor estimation result \(\hat{x}_k\) at each time \(k\). As illustrated in Fig. 4, we build the detector in the DT by comparing \(\hat{x}_k\) and \(\hat{x}_k\). The Chi-square detector generates a detection result \(q_k \in Q := \{0, 1, 2\}\) at time \(k\), where \(q_k = 0\) means the result is qualified, \(q_k = 1\) means the result is unqualified, \(q_k = 2\) means the result is detrimental. When \(q_k = 2\), the DT should always reject the estimation and send an alarm to the operators.

To design the detector, we define \(\phi_k := \hat{x}_k - \hat{x}_k\). Since \(\hat{x}_k\) and \(\hat{x}_k\) are Gaussian distributions, \(\phi_k\) is also a Gaussian distribution with a zero-mean vector and a covariance matrix, i.e., \(\phi \sim \mathcal{N}(0_{n_x}, \Sigma_\phi)\). Furthermore, we define that

\[
\chi^2_k := (\hat{x}_k - \hat{x}_k)^T \Sigma^{-1}_\phi (\hat{x}_k - \hat{x}_k).
\]

Then, \(\chi^2_k\) follows a Chi-square distribution. We define a Chi-square detector as the following:

\[
q_k = f_q(m_k) := \begin{cases} 
0, & \text{if } \chi^2_k \leq \rho_1; \\
1, & \text{if } \chi^2_k \in (\rho_1, \rho_2]; \\
2, & \text{if } \chi^2_k > \rho_2;
\end{cases}
\]

where \(\rho_1, \rho_2\) are two given thresholds, and they satisfy that \(\rho_2 > \rho_1 > 0\); \(f_q : \mathbb{R}^{n_x} \to Q\) is the detection function.

Using the above Chi-square detector, we can achieve fault detection. However, the work [16] has shown that intelligent attackers can constrain the ability of the Chi-square detector by manipulating the amount of injected data. In the following subsection, we will introduce a stealthy sensor attack that aims to remain stealthy while modifying the estimation.
E. General Setup of Signaling Game with Evidence

Due to the existence of attacks, the DT’s might not be able to monitor the actual value estimation directly. Instead, the DT’s can read a message provided by the estimator. According to our attack model, the attacker can compromise the estimator. Hence, the estimator can have two identities, i.e., a benign estimator or a malicious estimator. The DT aims to verify the estimator’s identity by monitoring the estimation results. As shown in Fig. 5, we can view DT’s monitoring process as a message-sending process, i.e., the estimator sends an estimation result to the DT for verification. To capture the interactions between the estimator and DT, we will formally define a Signaling Game with Evidence (SGE) as follows.

In an SGE, we have two players: one is the sender, and the other one is the receiver. The sender has a private identity, denoted by $\theta \in \Theta := \{\theta_0, \theta_1\}$, where $\theta_0$ means the sender is benign, and $\theta_1$ means the sender is malicious. According to its identity, the sender will choose a message $m \in \mathcal{M}$ and send it to the receiver. After observing the message, the receiver can choose an action $a \in \mathcal{A}$. Action $a = 1$ means that the receiver accepts the message, while $a = 0$ means the receiver rejects the message. The sender and receiver have their own utility functions $U_i : \Theta \times \mathcal{M} \times \mathcal{A} \rightarrow \mathbb{R}$, for $i \in \{s, r\}$. Fig. 6 illustrates how the data and information transmit in the proposed cyber-physical SGE.

Given mixed strategies $\sigma_s$ and $\sigma_r$, we define players’ expected utility functions as
\[
\hat{U}_s(\theta, \sigma_s, a) := \sum_{q \in \mathcal{Q}} \sigma_s(f_q(m)|\theta)U_s(\theta, m, a),
\]
\[
\hat{U}_r(\theta, m, \sigma_r) := \sum_{a \in \mathcal{A}} \sigma_r(a|f_q(m))U_r(\theta, m, a).
\]

To find the optimal mixed strategies of both players, we identify a Perfect Bayesian Nash Equilibrium (PBNE) of the SGE in the following definition.

**Definition 1:** A PBNE of a SGE is a strategy profile $(\sigma_s^*, \sigma_r^*)$ and a posterior belief $\pi(\theta)$ such that
\[
\sigma_s(f_q(m)|\theta) \in \arg\max_{\sigma_s \in \Gamma_s} \sum_{a \in \mathcal{A}} \sigma_r(a|f_q(m))\hat{U}_s(\theta, \sigma_s, a),
\]
\[
\sigma_r(a|f_q(m)) \in \arg\max_{\sigma_r \in \Gamma_r} \sum_{\theta \in \Theta} \pi(\theta)\hat{U}_r(\theta, m, \sigma_r),
\]
and the receiver updates the posterior belief using the Bayes’ rule, i.e.,
\[
\pi(\theta) = f_b(\pi'(\theta), f_q(m)) := \sum_{\theta' \in \Theta} \sigma_s(f_q(m)|\theta')\pi'(\theta)\pi(\theta').
\]

where $f_b : (0, 1) \times (0, 1) \rightarrow (0, 1)$ is the belief-update function, and $\pi'(\theta)$ is a prior belief of $\theta$.

**Remark 1:** Definition 1 identifies the optimal mixed strategies of the sender and receiver. One important thing is that at any PBNE, the belief $\pi(\theta)$ should be consistent with the optimal strategies, i.e., at the PBNE, belief $\pi(\theta)$ is independent of time $k$. Instead, $\pi(\theta)$ should only depend on detection results $q \in \mathcal{Q}$. Besides, we implement Bayes’s rule to deduce belief-update function (13).

In a SGE, there are different types of PBNE. We present three types of PBNE in the following definition.

**Definition 2:** (Types of PBNE) An SGE, defined by Definition 1 has three types of PBNE:

1) Pooling PBNE: The senders with different identities use identical strategies. Hence, the receiver cannot distinguish the identities of the sender based on the available evidence and message, i.e., the receiver will use the same strategies with the different senders.

2) Separated PBNE: Different senders will use different strategies based on their identities, and the receiver can distinguish the senders and use different strategies for different senders.

3) Partially-Separated PBNE: different senders will choose different, but not completely opposite strategies, i.e.,
\[
\sigma_s(f_q(m)|\theta_0) \neq 1 - \sigma_s(f_q(m)|\theta_1).
\]

**Remark 2:** In the separated PBNE, the receiver can obtain the identity of the senders by observing a finite sequence of evidence and messages. However, in the other two PBNE, the receiver may not be able to distinguish the senders’ identity.
Note that in real applications, the CPS will run the SGE repeatedly, and generate a sequence of detection results \( \mathcal{H}_k := \{q_0, q_1, \ldots, q_k\} \). At time \( k \), we define the posterior belief as

\[
\pi_k(\theta) = \Pr(\theta | \mathcal{H}_{k-1}).
\]

Whenever there is a new detection result \( q_k \), we can update the belief using \( \pi_{k+1}(\theta) = f_b(\pi_k(\theta), q_k) \), where function \( f_b \) is defined by (13). Belief \( \pi_{k+1}(\theta) \) will become a prior belief at time \( k + 1 \).

To this end, we will use the SGE framework to capture the interactions between the physical layer and the DT. We will find the optimal defense strategy of the DT by finding the PBNE. In the next section, we define the utility functions explicitly and find the PBNE of the proposed SGE. Given the PBNE, we can identify the optimal defense strategies.

### III. Equilibrium Results of the Cyber SGE

In this section, we aim to find the optimal defense strategy against a stealthy sensor attack. To this end, we first define the utility functions, which capture the profit earned by the players. Secondly, we identify the best response of the players when they observe or anticipate the other player’s strategy. Finally, we present a PBNE under the players’ best response and obtain an optimal defense strategy for the DT. We analyze the stability of the system under the stealthy attack.

#### A. SGE Setup for the CPSs

In this work, we use an SGE to capture the interactions between the physical estimator and the DT. In our scenario, the message set is just the estimation set, i.e., \( \mathcal{M} := \mathbb{R}^n. \) The DT monitors the estimation \( m_k \) and chooses an action \( a \in A := \{0, 1\} \). Action \( a = 1 \) means the estimation passes the verification, while action \( a = 0 \) means the verification fails, and the DT will send an alarm to the operators.

In the next step, we define the utility functions of both players, explicitly. Firstly, we define sender’s utility functions \( U_s(\theta, m, a) \). Since the sender has two identities, we need to define two types of utility functions for the sender. The utility function with \( \theta = 0 \) is defined by

\[
U_s(\theta_0, m_k, a_k) := -\mathbb{E}[(m_k - \hat{x}_k)^T (m_k - x_k)],
\]

In (14), we can see that \( U_s(\theta_0, m_k, a_k) \) is independent of action \( a_k \), and maximizing \( U_s(\theta_0, m_k, a_k) \) is equivalent to the estimation problem (6). Hence, given (14), the benign sender always sends the true estimation result \( \hat{x}_k \), defined by (7), regardless of action \( a_k \).

For the malicious estimator, we define its utility function as

\[
U_s(\theta_1, m_k, a_k) := \mathbb{E}[(m_k - \hat{x}_k)^T (m_k - x_k)] \cdot 1_{\{a_k=1\}},
\]

where \( 1_{\{s\}} = 1 \) if statement \( s \) is true. In (15), we see that the motivation of the attacker is to deviate the system state as much as possible while remaining undiscovered. However, the attacker’s utility will be zero if the DT detects the attack.

Secondly, we define the DT’s utility function. Note that the DT’s utility function should depend on the identity of the sender. When the estimator is benign, i.e., \( \theta = 0 \), the DT should choose \( a_k = 1 \) to accept the estimation. When the estimator is malicious, i.e., \( (\theta = 1) \), the DT should choose \( a_k = 0 \) to reject the estimation and send an alarm to the operators. Given the motivations, we define \( U_r(\theta, \hat{x}, a) \)

\[
U_r(\theta_0, m_k, a_k) := -(\hat{x}_k - m_k)^T \hat{Q}_0 (\hat{x}_k - m_k) \cdot 1_{\{a_k=0\}},
\]

\[
U_r(\theta_1, m_k, a_k) := -(\hat{x}_k - m_k)^T \hat{Q}_1 (\hat{x}_k - m_k) \cdot 1_{\{a_k=1\}},
\]

where \( \hat{Q}_0, \hat{Q}_1 \in \mathbb{R}^{n_x \times n_x} \) are positive-definite matrices. The weighting matrices will affect the receiver’s defense strategy. A large value of \( \text{tr}(\hat{Q}_1) \) will lead to a conservative strategy, while a large value of \( \text{tr}(\hat{Q}_0) \) will lead a radical one. Readers can receive more details in Proposition 1.

In the next subsection, we analyze the behaviors of the players and obtain the best-response strategies. Note that function \( U_r(\theta, m_k, a_k) \) is deterministic. The reason is that the DT can observe \( \hat{x}_k \) and \( \hat{x}_k \) at time \( k \), explicitly. However, the physical estimator cannot observe \( x_k \) at time \( k \).

#### B. Best Response of the Players and a PBNE of the SGE

We first analyze the best response of the DT. Given belief \( \pi_k(\theta) \), message \( m_k \), and detection result \( q_k \), we present the following theorem to identify DT’s best response.

**Proposition 1:** (DT’s Best Response) Given

\[
\sigma_r^*(a_k | q_k) = \begin{cases} 1, & \text{if } q_k \neq 2, \pi_k(\theta_0) \geq \beta; \\ 0, & \text{if } q_k \neq 2, \pi_k(\theta_0) < \beta; \\ 0, & \text{if } q_k = 2; \end{cases}
\]

where \( \beta \) is defined by

\[
\beta := \frac{\langle \hat{x}_k - m_k, (\hat{x}_k - m_k)^T (\hat{Q}_0 + \hat{Q}_1)(\hat{x}_k - m_k) \rangle}{\langle \hat{x}_k - m_k, (\hat{x}_k - m_k)^T \hat{Q}_1 (\hat{x}_k - m_k) \rangle}.
\]

**Proof:** Note that

\[
\mathbb{E}[U_r(\theta, m_k, a_k = 1)|q_k] \geq \mathbb{E}[U_r(\theta, m_k, a_k = 0)|q_k]
\]

\[
\Leftrightarrow a_k = 1 \text{ if } \pi_k(\theta_0) \geq \beta;
\]

where \( \beta \) is defined by (15). This completes the proof.

**Remark 3:** Given Proposition 1, we note that the DT uses a pure strategy since it can make its decision after observing detection result \( q_k \) and message \( m_k \).

In the next step, we consider the best response of the estimator. If the estimator is benign, i.e., \( \theta = \theta_0 \), the optimal estimation should be (7). Therefore, the optimal utility of the benign estimator is given by

\[
U_s(\theta_0, \hat{x}_k, a_k) = \mathbb{E}[(\hat{x}_k - x_k)^T (\hat{x}_k - x_k)] = \text{tr}(\hat{P}_k),
\]

where \( \text{tr}(P) \) is the trace of matrix \( P \). The following theorem shows the optimal mixed strategy of the benign estimator.

**Proposition 2:** (Best Response of the Benign Estimator) Given the DT’s best response (16), the optimal mixed strategy of the benign estimator, i.e., \( \theta = \theta_0 \), is given by

\[
\sigma_s^*(f_q(m_k) = 0 | \theta_0) = F_\chi (\rho_1, n_x),
\]

\[
\sigma_s^*(f_q(m_k) = 1 | \theta_0) = F_\chi (\rho_2, n_x) - F_\chi (\rho_1, n_x),
\]

\[
\sigma_s^*(f_q(m_k) = 2 | \theta_0) = 1 - F_\chi (\rho_2, n_x),
\]
where \( \hat{x}_k \) is defined by (7), \( F_X(\rho, n) : \mathbb{R}_+ \rightarrow [0, 1] \) is the Cumulative Distribution Function (CDF) of the Chi-square distribution with \( n \in \mathbb{Z}_+ \) degrees.

**Proof:** Note that the benign estimator will choose \( m_k = \hat{x}_k \), defined by (7). According to definition (11), we know that \((\hat{x}_k - \hat{x}_k)\) follows a Chi-square distribution with \( n_x \) degrees. Hence, we have

\[
\Pr(\hat{x}_k \leq \rho_1) = F_X(\rho_1, n_x), \\
\Pr(\hat{x}_k > \rho_2) = 1 - F_X(\rho_2, n_x), \\
\Pr(\hat{x}_k \in (\rho_1, \rho_2]) = F_X(\rho_2, n_x) - F_X(\rho_1, n_x).
\]

Combining the above equations with Chi-square detector (12) yields mixed strategies (19)-(21).

**Remark 4:** Note that the benign estimator always choose the optimal estimation \( \xi \). However, from DT’s perspective in this game, the real mixed strategies of the benign estimator are (19)-(21) because of uncertainty introduced by the noises.

From the perspective of the malicious estimator, it needs to select \( \sigma_s(f_x(m_k)\theta_1) \) such that \( \pi_k(\theta_0) \geq \beta \). Given the attackers’ incentive, we obtain the following theorem.

**Proposition 3:** (Best Response of the Malicious Estimator)

\[
\sigma_s^*(f_x(f_{\xi,k,1} = 0|\theta_1) = F_X(\rho_1, n_x), \\
\sigma_s^*(f_x(f_{\xi,k,2} = 1|\theta_1) = 1 - F_X(\rho_1, n_x),
\]

where \( \xi_{k,1}, \xi_{k,2} \) are the solutions to the following problems:

\[
\xi_{k,1} \in \arg \max_{m \in \mathcal{M}_{p_1}(\hat{x}_k)} U_\theta(\xi_{k,1}, m, \alpha_k = 1), \\
\xi_{k,2} \in \arg \max_{m \in \mathcal{M}_{p_2}(\hat{x}_k)} U_\theta(\xi_{k,2}, m, \alpha_k = 1),
\]

with spaces \( \mathcal{M}_{p_1}(\hat{x}_k) \) and \( \mathcal{M}_{p_2}(\hat{x}_k) \) defined by

\[
\mathcal{M}_{p_1}(\hat{x}_k) := \{ m \in \mathbb{R}^n : \|\hat{x}_k - m\|_{\Sigma_0^{-1}}^2 \leq \rho_1 \}, \\
\mathcal{M}_{p_2}(\hat{x}_k) := \{ m \in \mathbb{R}^n : \|\hat{x}_k - m\|_{\Sigma_0^{-1}}^2 \in (\rho_1, \rho_2] \}.
\]

**Proof:** Firstly, the attacker has no incentive to choose \( m_k \notin \mathcal{M}_{p_1}(\hat{x}_k) \cup \mathcal{M}_{p_2}(\hat{x}_k) \) because its utility will be zero. Secondly, the attacker aims to choose the mixed strategy \( \sigma_s^*(q_k = 1|\theta_1) \) as large as possible since it can make a higher damage to the system. Then, we show that the optimal mixed strategy of the attacker is given by (22) and (23). To do this, based on (13), we consider the following belief update:

\[
\pi_k+1(\theta_0) = \sigma_s^*(q_k \neq 0|\theta_0)\pi_k(\theta_0) + \sigma_s(0|\theta_0)(1 - \pi_k(\theta_0)) = \Delta \pi_k(\theta_0) + \sigma_s(q_k \neq 1|\theta_1),
\]

where \( \Delta \pi_k(\theta_0) \) is defined by

\[
\Delta \pi_k(q_k \neq 0) := \sigma_s^*(q_k \neq 0|\theta_0) - \sigma_s(q_k \neq 0|\theta_1).
\]

Rearranging (26) yields that

\[
\frac{\pi_{k+1}(\theta_0)}{\pi_k(\theta_0)} = \frac{\sigma_s^*(q_k \neq 0|\theta_0)\pi_k(\theta_0)}{\Delta \pi_k(\theta_0) + \sigma_s(q_k \neq 0|\theta_1)}.
\]

Given that \( \pi_k(\theta_0) \in (0, 1) \), we have

\[
\frac{\pi_{k+1}(\theta_0)}{\pi_k(\theta_0)} > 1, \text{ if } \Delta \pi_k(q_k \neq 0) > 0; \\
\frac{\pi_{k+1}(\theta_0)}{\pi_k(\theta_0)} = 1, \text{ if } \Delta \pi_k(q_k \neq 0) = 0; \\
\frac{\pi_{k+1}(\theta_0)}{\pi_k(\theta_0)} < 1, \text{ if } \Delta \pi_k(q_k \neq 0) < 0.
\]

When \( \pi_k(\theta_0) \in [\beta, 1) \), the attacker has to choose \( \Delta \pi_k(q_k \neq 0) = 0 \) to maintain the belief at a constant. Otherwise, the belief will decrease continuously. When the belief \( \pi_k(\theta_0) \) stays lower than \( \beta \), the DT will send an alert to the operators. Hence, the optimal mixed strategies of the malicious estimator are given by (22) and (23).

Given the results of Propositions 1, 2, and 3, we present the following theorem to characterize a unique pooling PBNE.

**Theorem 1:** (The PBNE of the Proposed SGE) The proposed cyber SGE has a unique pooling PBNE. At the PBNE, the optimal mixed strategies of the benign and malicious sender are presented by (19)-(21) and (22)-(23). The DT has a pure strategy defined by (16). At the PBNE, belief \( \pi_k(\theta_0) \in [\beta, 1) \) is a fixed point of function \( f_b \), i.e.,

\[
\pi_k^*(\theta_0) = f_b(\pi_k^*(\theta_0), q_k), \text{ for } q_k \in \mathcal{Q}.
\]

**Proof:** We first show the existence of the pooling PBNE. Suppose that both estimators use strategies (19)-(21), (22)-(23), respectively, and the DT uses (16). Then, no player has incentive to move since these are already the optimal strategies. Besides, for any \( \theta \in \Theta, q_k \in \mathcal{Q} \), we note that

\[
\pi_{k+1}(\theta) = f_b(\pi_k^*(\theta), q_k) = \pi_k^*(\theta),
\]

where \( f_b \) is defined by (13). Hence, \( \pi_k^*(\theta) \) is a fixed point of function \( f_b \), and the belief remain at \( \pi_k^*(\theta) \), which means the belief stays consistently with the optimal strategies of the sender and receiver. Hence, the proposed strategies pair \( (\pi_k^*, \pi_b^*) \) is a PBNE.

Secondly, we show that pooling PBNE is unique. We note that the DT and benign estimator have no incentive to move since they already choose their best strategies. In Proposition 3 we already show that the attacker cannot change its mixed strategies. Otherwise, the belief cannot remain constant. Hence, pooling PBNE is unique.

**Remark 5:** Theorem 1 shows that the SGE admits a unique pooling PBNE, which means that an intelligent attacker can use its stealthy strategies to avoid being detected by the DT.

In the next subsection, we will analyze the stability of the system under the stealthy attack. Besides, we will also evaluate the loss caused by the attack.

### C. Estimated Loss Under the Stealthy Attack

In the previous subsection, we have shown the PBNE in which the attacker can use a stealthy strategy to pass the verification of the DT. In this subsection, we will quantify the loss under the attack. Before presenting the results, we need the following lemma.

**Lemma 2:** Given \( \rho_i, \xi_{k,i} \) for \( i \in \{1, 2\} \), we have the following relationship:

\[
\rho_i \lambda_{\max}(\Sigma_{\rho_i}) \geq (\hat{x}_k - \xi_{k,i})^T(\hat{x}_k - \xi_{k,i}), \text{ for } i = 1, 2,
\]
where $\lambda_{\text{max}}(\Sigma)$ is the greatest eigenvalue of matrix $\Sigma$.

Proof: Firstly, we note that $U_s$ is strictly convex in $m_k$. The solution to problem (24) and (25) must stay at the boundary. Hence, we have

$$\rho_i = (\bar{x}_k - \xi_{k,i})^T \Sigma^{-1}_\phi (\bar{x}_k - \xi_{k,i})$$

$$\geq \frac{(\bar{x}_k - \xi_{k,i})^T (\bar{x}_k - \xi_{k,i})}{\lambda_{\text{max}}(\Sigma_\phi)}$$

(27)

Rearranging (27) yields that

$$\rho_i \lambda_{\text{max}}(\Sigma_\phi) \geq (\bar{x}_k - \xi_{k,i})^T (\bar{x}_k - \xi_{k,i}).$$

This completes the proof.

Considering different estimators, we define two physical cost functions $J_0$ and $J_1$, i.e.,

$$J_0 := \lim_{N \to \infty} \mathbb{E} \left\{ \frac{1}{N} \sum_{k=0}^{N-1} x_k^T Q x_k + \mu^T(m_k) R \mu(m_k) \right\} \theta_0,$$

$$J_1 := \lim_{N \to \infty} \mathbb{E} \left\{ \frac{1}{N} \sum_{k=0}^{N-1} x_k^T Q x_k + \mu^T(m_k) R \mu(m_k) \right\} \theta_1.$$

We define a loss function $\Delta J := J_1 - J_0$ to quantify the loss caused by the stealthy sensor attack. Given pooling PBNE defined by Theorem 1 we provide an upper-bound of $\Delta J$ in the following theorem.

Theorem 2: (Bounded Loss) The proposed framework can guarantee stability of the CPSSs, and the value of function $\Delta J$ is bounded by a constant, i.e.,

$$\Delta J = J_1 - J_0 \leq \alpha_1 \text{tr}(\hat{P}^*) - \alpha_0 \text{tr}(\hat{P}^*) + \alpha_1 \rho_1 \lambda_{\text{max}}(\Sigma_\phi) F_X(\rho_1, n_x) + \alpha_1 \rho_2 \lambda_{\text{max}}(\Sigma_\phi) (1 - F_X(\rho_1, n_x)),$$

where $\alpha_0$ and $\alpha_1$ are defined by

$$\alpha_0 := \frac{\lambda_{\text{min}}(R_K) (\lambda_{\text{min}}(G) - \lambda_{\text{min}}(R_K))}{\lambda_{\text{min}}(G)},$$

$$\alpha_1 := \frac{\lambda_{\text{max}}(R_K) (\lambda_{\text{max}}(R_K) + 2 \lambda_{\text{max}}(G))}{\lambda_{\text{max}}(G)},$$

and $R_K := K^T R K$, $G := Q + R K^T \lambda_{\text{max}}(W)$ and $\lambda_{\text{min}}(W)$ are the greatest and smallest eigenvalues of matrix $W$.

Proof: Firstly, we note that

$$\mathbb{E} \left[ \|x_k\|_{\hat{Q}}^2 + \|K \hat{x}_k\|_R^2 \right]$$

$$= \mathbb{E} \left[ \|x_k\|_{\hat{Q}}^2 + 2 x_k^T R K \hat{e}_k + \|\hat{e}_k\|_{R_K}^2 \right]$$

$$\geq \mathbb{E} \left[ \sqrt{\lambda_{\text{min}}(G)} x_k^T + \frac{\lambda_{\text{min}}(R_K)}{\sqrt{\lambda_{\text{min}}(G)}} \|\hat{e}_k\| \right]^2$$

$$+ \frac{\lambda_{\text{min}}(R_K) (\lambda_{\text{min}}(G) - \lambda_{\text{min}}(R_K))}{\lambda_{\text{min}}(G)} \|\hat{e}_k\|^2 \right] \geq \alpha_0 \text{tr}(\hat{P}_k),$$

Using the above inequality, we observe that

$$J_0 = \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} \mathbb{E} \left[ \|x_k\|_{\hat{Q}}^2 + \|K \hat{x}_k\|_R^2 \right]$$

$$\geq \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} \alpha_0 \text{tr}(\hat{P}_k) = \alpha_0 \text{tr}(\hat{P}^*),$$

(29)

where $\hat{P}^*$ is defined by (8). Secondly, we also note that

$$\mathbb{E} \left[ x_k^T Q x_k + \mu^T(\xi_{k,i}) R \mu(\xi_{k,i}) \right]$$

$$= \mathbb{E} \left[ \|x_k\|_{\tilde{Q}}^2 + \|x_k + \hat{e}_k + \xi_{k,i} - \bar{x}_k\|_{R_K}^2 \right]$$

$$\leq \mathbb{E} \left[ \|x_k\|_{\tilde{Q}}^2 + 2 x_k^T R_K \hat{e}_k + 2 x_k^T R_K (\xi_{k,i} - \bar{x}_k) \right]$$

$$+ 2 \hat{e}_k^T R_K (\xi_{k,i} - \bar{x}_k) + \|\xi_{k,i} - \bar{x}_k\|_{R_K}^2 + \|\hat{e}_k\|_{R_K}^2$$

$$\leq \mathbb{E} \left[ 3 \|x_k\|_{\tilde{Q}}^2 + \frac{\lambda_{\text{max}}(R_K)}{\lambda_{\text{max}}(G)} \left( \|\hat{e}_k\|^2 + \|\xi_{k,i} - \bar{x}_k\|^2 \right) \right]$$

$$+ 2 \lambda_{\text{max}}(R_K) \|\xi_{k,i} - \bar{x}_k\|^2 \right]$$

$$\leq 3 \|x_k\|_{\tilde{Q}}^2 + \alpha_1 \text{tr}(\hat{P}_k) + \alpha_1 \rho_1 \lambda_{\text{max}}(\Sigma_\phi),$$

(30)

We complete the squares to deduce the second inequality of (30). Similarly, we have

$$J_1 = \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} \left( F_X(\rho_1, n_x) \mathbb{E} \left[ \|x_k\|_{\tilde{Q}}^2 + \|\xi_{k,i}\|_{R_K}^2 \right] \right)$$

$$+ \left( 1 - F_X(\rho_1, n_x) \right) \mathbb{E} \left[ \|x_k\|_{\tilde{Q}}^2 + \|\xi_{k,i}\|_{R_K}^2 \right]$$

$$\leq \lim_{N \to \infty} \frac{3}{N} \sum_{k=0}^{N-1} \|x_k\|_{\tilde{Q}}^2 + \alpha_1 \rho_1 \lambda_{\text{max}}(\Sigma_\phi) F_X(\rho_1, n_x)$$

$$+ \alpha_1 \rho_2 \lambda_{\text{max}}(\Sigma_\phi) (1 - F_X(\rho_1, n_x)) + \alpha_1 \text{tr}(\hat{P}^*),$$

(31)

where $\hat{P}^*$ is defined by (10). Combining inequalities (29) and (31) yields inequality (28). Hence, the system is stable, and the impact of the attack is bounded by a constant.

Remark 6: Theorem 2 shows that the difference between $J_0$ and $J_1$ is bounded, i.e., the stealthy estimation attack cannot deviate the system to an arbitrary point even if the attacker has an infinite amount of time.

In the next subsection, we will use an application to evaluate the performance of the proposed defense strategies.

IV. SIMULATION RESULTS

In this section, we use a two-link Robotic Manipulator (RM) to investigate the impact of the estimation attacks. In the experiments, we use different case studies to analyze the performance of the proposed defense framework.
A. Experimental Setup

Fig. 7 illustrates the physical structure of the two-link RM. Variables $g_1$ and $g_2$ are the angular positions of Links 1 and 2. We summarize the parameters of the RM in Table II.

| Parameter | Description | Value |
|-----------|-------------|-------|
| $l_1$     | Length of Link 1 | 0.6 m |
| $l_2$     | Length of Link 2 | 0.4 m |
| $r_1$     | Half Length of Link 1 | 0.3 m |
| $r_2$     | Half Length of Link 2 | 0.2 m |
| $m_1$     | Mass of Link 1 | 6.0 kg |
| $m_2$     | Mass of Link 2 | 4.0 kg |
| $I_1$     | Inertia of Link 1 on z-axis | 1 kg-m$^2$ |
| $I_2$     | Inertia of Link 2 on z-axis | 1 kg-m$^2$ |

Let $g = [g_1, g_2]^T$ be the angular vector and $\tau = [r_1, r_2]^T$ be the torque input. According to the Euler-Lagrange Equation, we obtain the dynamics of the two-link RM as

$$M(g)\ddot{g} + S(g, \dot{g}) = \tau,$$  \hspace{1cm} (32)

where matrices $M(g)$ and $S(g, \dot{g})$ are defined by

$$M(g) := \begin{bmatrix} a + b \cos(g_2) & \delta + b \cos(g_2) \\ \delta + b \cos(g_2) & \delta \end{bmatrix},$$

$$S(g, \dot{g}) := \begin{bmatrix} -b \sin(g_2) \dot{g}_2 & -b \sin(g_2) (\dot{g}_1 + \dot{g}_2) \\ b \sin(g_2) \dot{g}_1 & \dot{g}_1 \end{bmatrix},$$

$$a := I_1 + I_2 + \eta_1 r_1^2 + \eta_2 (l_1^2 + l_2^2),$$

$$b := \eta_2 l_1 r_2, \quad \delta := I_2 + \eta_2 r_2^2.$$  \hspace{1cm} (33)

To control the two-link RM, we let $\tau := M(g)a_g + S(g, \dot{g})\ddot{g}$, where $a_g \in \mathbb{R}^2$ is the acceleration that we need to design. Note that $M(g)$ is positive-definite, i.e., $M(g)$ is invertible. Hence, substituting $\tau$ into (32) yields that

$$M(g)\ddot{g} = M(g)a_g \Rightarrow \ddot{g} = a_g.$$  \hspace{1cm} (34)

Let $p \in \mathbb{R}^2$ be the position of RM’s end-effector. We have

$$\ddot{p} = H(g)\ddot{g} + \dot{H}(g)\dot{g} = H(g)a_g + \dot{H}(g)\dot{g},$$  \hspace{1cm} (35)

where $H(g)$ is the Jacobian matrix. Then, we substitute $a_g := H^{-1}(g)(u - \dot{H}(g))$ into (33), arriving at $\ddot{p} = u$. Let $x = [p^T, \dot{p}^T]^T$ be the continuous-time state. Then, we obtain a continuous-time linear system $\dot{x} = A_tx + B_tu$. Given a sampling time $\Delta T > 0$, we discretize the continuous-time system to obtain system model $\{1\}$. We let $y_k$ and $z_k$ be

$$y_k = x_k + v_k, \quad z_k = p(k\Delta T) + d_k$$  \hspace{1cm} (36)

We assume that the DT uses security-protected cameras to identify the position of the end-effector.

In the experiments, we let the RM to draw a half circle on a two-dimensional space. The critical parameters are given by

$$\beta = 0.65, \quad n_x = 4, \quad \rho_1 = 9.49, \quad \rho_2 = 18.47, \quad F_\text{x}(\rho_1, n_x) = 0.95, \quad F_\text{y}(\rho_2, n_x) = 0.999.$$  \hspace{1cm} (37)

We have three case studies: a no-attack case, a normal-attack case, and a stealthy-attack case. In the normal-attack case, the attacker is not aware of the defense strategies and deviate the system from the trajectory, directly. In the last case, the attacker aims to tamper the estimation without being detected.

Figures 8, 9, and 10 illustrate the simulation results of the case studies. In Fig. 8 (a), we can see that the RM can track the trajectory smoothly when there is no attack. However, we note that DT’s estimation is worse than the physical estimation, which coincides with our expectation. Figures 8 (b) and (c) show the value of the Chi-square and the belief of the DT. In the no-attack case, the Chi-square detector will remain silent with a low false alarm rate, and the belief stays at a high level. In Fig. 9 (a), the attackers deviate the system without considering the detection. Even though DT’s estimation is not accurate, the attacker cannot tamper that. Therefore, the detector will rapidly locate the attack and send alarms to the operators. The belief of $\theta_0$ will remain at the bottom line. In Fig. 10, differently, the stealthy attackers know the defense strategies and try to maintain the Chi-square value below threshold $\rho_1$. However, the behavior mitigates the impact of the attack, which also coincides with the result of Theorem 2.
Figure 11 illustrates the Mean Square Errors (MSE) of different cases. Figure 11 (a) presents that the MSE of the physical estimator is much smaller than the DT’s estimator, i.e., the physical estimator can provide more accurate sensing information. However, in Figure 11 (b), we can see that the attacker can deviate the physical estimation to a significant MSE. Besides, under the DT’s supervision, the stealthy attacker fails to generate a large MSE. The above results show that the proposed defense mechanism succeeds in mitigating the stealthy attacker’s impact.

### V. Conclusions

In this paper, we have considered a stealthy estimation attack, where an attack can modify the estimation results to deviate the system without being detected. To mitigate the impact of the attack on physical performance, we have developed a Chi-square detector, running in a Digital Twin (DT). The Chi-square detector can collect DT’s observations and the physical estimation to verify the identity of the estimator. We have used a Signaling Game with Evidence (SGE) to study the optimal attack and defense strategies. Our analytical results have shown that the proposed framework can constrain the attackers’ ability and guarantee the stability.

### References

[1] K.-D. Kim and P. R. Kumar, “Cyber–physical systems: A perspective at the centennial,” *Proceedings of the IEEE*, vol. 100, no. Special Centennial Issue, pp. 1297–1308, 2012.

[2] L. Wang, M. Tönngrn, and M. Onori, “Current status and advancement of cyber-physical systems in manufacturing,” *Journal of Manufacturing Systems*, vol. 37, pp. 517–527, 2015.

[3] Y. Mo, T. Holtz, D. P. Shepard, J. A. Bhatti, and T. E. Humphreys, “Drone hack: Spoofing attack demonstration on a civilian unmanned aerial vehicle,” 2012.

[4] Y. Mo and B. Sinopoli, “Integrity attacks on cyber-physical systems,” in *Proceedings of the 1st international conference on High Confidence Networked Systems*. ACM, 2012, pp. 47–54.

[5] F. Pasqualetti, F. Dorfler, and F. Bullo, “Control-theoretic methods for cyber-physical security: Geometric principles for optimal cross-layer resilient control systems,” *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 110–127, 2015.

[6] D. Ding, Q.-L. Han, Y. Xiang, X. Ge, and X.-M. Zhang, “A survey on security control and attack detection for industrial cyber-physical systems,” *Neurocomputing*, vol. 275, pp. 1674–1683, 2018.

[7] F. Pasqualetti, F. Dorfler, and F. Bullo, “Attack detection and identification in cyber-physical systems,” *IEEE transactions on automatic control*, vol. 58, no. 11, pp. 2715–2729, 2013.

[8] F. Pasqualetti, F. Dorfler, and F. Bullo, “Attack detection and identification in cyber-physical systems,” *IEEE transactions on automatic control*, vol. 61, no. 9, pp. 2618–2624, 2015.

[9] Y. Chen, S. Kar, and J. M. Moura, “Optimal attack strategies subject to detection constraints against cyber-physical systems,” *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 1157–1168, 2017.

[10] Z. Xu and Q. Zhu, “Cross-layer secure and resilient control of delay-sensitive networked robot operating systems,” in *2015 IEEE Conference on Control Technology and Applications (CCTA)*. IEEE, 2018, pp. 1712–1717.

[11] W. Luo, T. Hu, C. Zhang, and Y. Wei, “Digital twin for cnc machine tool: modeling and using strategy,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 3, pp. 1129–1140, 2019.

[12] S. Shen, Y. Li, H. Xu, and Q. Cao, “Digital twin-driven product design, manufacturing and service with big data,” *The International Journal of Advanced Manufacturing Technology*, vol. 94, no. 9-12, pp. 3563–3576, 2018.

[13] B. Bielefeldt, J. Hochhalter, and D. Hartl, “Computationally efficient analysis of sma sensory particles embedded in complex aerostuctures using a substructure approach,” in *ASME 2015 Conf. on Smart Materials, Adaptive Structures and Intelligent Systems*. American Society of Mechanical Engineers Digital Collection, 2016.

[14] M. H. Manshaei, Q. Zhu, T. Alpcan, T. Başar, and J.-P. Hubaux, “Game theory meets network security and privacy,” *ACM Computing Surveys (CSUR)*, vol. 45, no. 3, pp. 25, 2013.

[15] F. Pasqualetti, F. Dorfler, and F. Bullo, “Control-theoretic methods for cyber-physical security: Geometric principles for optimal cross-layer resilient control systems,” *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 82–92, 2016.

[16] T. Hu, C. Zhang, and Y. Wei, “Digital twin for cnc machine tool: modeling and using strategy,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 3, pp. 1129–1140, 2019.

[17] M. Pajic, I. Lee, and G. J. Pappas, “Attack-resilient state estimation for cyberphysical systems under adversarial attacks,” *IEEE Transactions on Industrial Control Systems*, vol. 30, p. 62, 2014.

[18] F. Pasqualetti, F. Dorfler, and F. Bullo, “Control-theoretic methods for cyber-physical security: Geometric principles for optimal cross-layer resilient control systems,” *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 110–127, 2015.

[19] H. Fawzi, P. Tabuada, and S. Diggavi, “Secure estimation and control for noisy dynamical systems,” *IEEE transactions on Control of Network Systems*, vol. 4, no. 1, pp. 82–92, 2016.

[20] Y. H. Chang, Q. Hu, and C. J. Tomlin, “Secure estimation based kalman filter for cyber–physical systems against sensor attacks,” *Automatica*, vol. 95, pp. 399–412, 2018.

[21] D. Ding, Q.-L. Han, Y. Xiang, X. Ge, and X.-M. Zhang, “A survey on security control and attack detection for industrial cyber-physical systems,” *Neurocomputing*, vol. 275, pp. 1674–1683, 2018.

[22] F. Pasqualetti, F. Dorfler, and F. Bullo, “Attack detection and identification in cyber-physical systems,” *IEEE transactions on automatic control*, vol. 58, no. 11, pp. 2715–2729, 2013.

[23] Y. Mo and B. Sinopoli, “On the performance degradation of cyber-physical systems under stealthy integrity attacks,” *IEEE Transactions on Automatic Control*, vol. 61, no. 9, pp. 2618–2624, 2015.

[24] Y. Chen, S. Kar, and J. M. Moura, “Optimal attack strategies subject to detection constraints against cyber-physical systems,” *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 1157–1168, 2017.

[25] Z. Xu and Q. Zhu, “Cross-layer secure and resilient control of delay-sensitive networked robot operating systems,” in *2015 IEEE Conference on Control Technology and Applications (CCTA)*. IEEE, 2018, pp. 1712–1717.

[26] W. Luo, T. Hu, C. Zhang, and Y. Wei, “Digital twin for cnc machine tool: modeling and using strategy,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 3, pp. 1129–1140, 2019.

[27] F. Tao, J. Cheng, Q. Qi, M. Zhang, H. Zhang, and F. Sui, “Digital twin-driven product design, manufacturing and service with big data,” *The International Journal of Advanced Manufacturing Technology*, vol. 94, no. 9-12, pp. 3563–3576, 2018.

[28] B. Bielefeldt, J. Hochhalter, and D. Hartl, “Computationally efficient analysis of sma sensory particles embedded in complex aerostuctures using a substructure approach,” in *ASME 2015 Conf. on Smart Materials, Adaptive Structures and Intelligent Systems*. American Society of Mechanical Engineers Digital Collection, 2016.

[29] M. H. Manshaei, Q. Zhu, T. Alpcan, T. Başar, and J.-P. Hubaux, “Game theory meets network security and privacy,” *ACM Computing Surveys (CSUR)*, vol. 45, no. 3, pp. 25, 2013.

[30] S. Shen, Y. Li, H. Xu, and Q. Cao, “Signaling game based strategy of intrusion detection in wireless sensor networks,” *Computers & Mathematics with Applications*, vol. 62, no. 6, pp. 2404–2416, 2011.

[31] J. Pawlick, E. Colbert, and Q. Zhu, “Modeling and analysis of leaky deception using signaling games with evidence,” *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 7, pp. 1871–1886, 2018.

[32] A. Papoulis and S. U. Pillai, *Probability, random variables, and stochastic processes*. Tata McGraw-Hill Education, 2002.

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

[34] 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 (TISSEC)*, vol. 14, no. 1, p. 13, 2011.

[35] C. Tankard, “Advanced persistent threats and how to monitor and deter them,” *Network security*, vol. 2011, no. 8, pp. 16–19, 2011.

[36] B. I. Cohen, K. Kilian, and P. Rogaway, “The security of the cipher block chaining message authentication code,” *Journal of Computer and System Sciences*, vol. 61, no. 3, pp. 362–390, 2000.

[37] R. C. Merkle, “A certified digital signature,” in *Conference on the Theory and Application of Cryptology*. Springer, 1989, pp. 218–238.