On Distinguishability of Anomalies as Physical Faults or Actuation Cyberattacks

Increased automation has created an impetus to integrate infrastructure with wide-spread connectivity in order to improve efficiency, sustainability, autonomy, and security. Nonetheless, this reliance on connectivity and the inevitability of complexity in this system increase the vulnerabilities to physical faults or degradation and external cyber-threats. However, strategies to counteract faults and cyberattacks would be widely different and thus it is vital to not only detect but also to identify the nature of the anomaly that is present in these systems. In this work, we propose a mathematical framework to distinguish between physical faults and cyberattack using a sliding mode based unknown input observer. Finally, we present simulation case studies to distinguish between physical faults and cyberattacks using the proposed distinguishability metric and criterion. The simulation results show that the proposed framework successfully distinguishes between faults and cyberattacks. [DOI: 10.1115/1.4065292]

Keywords: fault detection and diagnosis, distinguishability metric, fault versus cyberattack

1 Introduction

The growing need for efficiency, coordination, precision, and autonomy has led to the integration of cybernetic components with physical infrastructure through information and communication technologies (ICT). Such physical systems with embedded networks of sensors, actuators, controllers are commonly described as cyber-physical systems (CPS). Currently, such CPS has garnered a lot of interest in the areas of smart grid [1], manufacturing [2], mobility [3], and many others. Thus, for reliable operation of these safety-critical systems, ensuring safety and security of these systems against faults and cyberattacks has become obligatory.

1.1 Motivation. The impact of faults and cyberattacks on CPS may be disparate [4]. On one hand, faults may arise due to natural degradation of system components or physical abuse. On the other hand, cyberattack is specically crafted by an adversary to drive system toward unintended states while evading detection by the system administrator. The wide-range of possibilities for physical faults and cyberattacks also make it challenging to distinguish between them from system measurements. Particularly, faults can be incipient or rapidly evolving leading to runway e cts [5]. In contrast, some cyberattacks can be passive (such as eavesdropping attack) or stealthy or can deny services from the system altogether. Additionally, the adversary can also design cyberattacks such that it can mimic behavior of faults in systems [6] or coordinate series of multiple faults in the systems [7]. More importantly, if faults and cyberattacks are wrongly classified, they may lead to incorrect remedial actions and cause severe disruptions.

1.2 Literature Review. Even though detection and isolation of both faults and cyberattacks have been a field of active research over the last decade, e orts to distinguish them has remained under-explored. In a distributed sensor network, hidden markov models have been used to distinguish between faulty and malicious data [8]. On the other hand, in Ref. [6], cyberattacks that maliciously trip relays to disrupt power distribution have been distinguished from faults by observing the flow of fault current in the power grid. The first e ort toward formalizing attack policies began with the introduction of an attack-space representation with respect to adversary’s system knowledge, disclosure, and disruption resources [4]. This work provides replay, zero-dynamics, and bias-injection attack policies. They also present stealthy bias-injection attack policy under incomplete system knowledge. Under a multi-agent scenario, Ref. [9] proposed an $H_\infty$ optimization-based observer design that distinguishes between in-domain faults and false-data injection attacks to the sensor measurement. The formulation considered in this work is restrictive in the sense that cyberattacks only act pair-wise agents while the faults act all the agents in the system. In contrast, Ref. [10] uses both physical and cyber properties of a multi-agent system (speci cally smart grid) in order to achieve the same. Additionally, data-driven strategies to distinguish between faults and cyberattacks have been tackled in the context of smart grids in Refs. [11,12] and for smart buildings in Refs. [13,14]. Finally, Ref. [15] utilized both model-based detection strategies along with information technology solutions to achieve the same.

1.3 Research Gap and Contribution. Literature in fault diagnostics and cyber-security reveal that a mathematical framework for distinguishing faults and cyberattacks for linear system has not been proposed, to the best of our knowledge. Thus, to address this gap we use a sliding mode observer to estimate an anomalous input to the system and provide a criterion to distinguish whether the anomalous
input is a fault or a cyberattack. Although estimation of anomalous input is often implemented by a Luenberger unknown input observers [16], the linear output injection term often causes issues of convergence therein. In contrast, the nonlinear switching laws of the designed sliding mode-based observer offer better accuracy and finite-time convergence.

1.4 Organization of the Paper. The remainder of the paper is organized as follows: Sec. 2 describes the problem setup, Sec. 3 presents the distinguishability criterion, Sec. 4 shows the validation of our framework through simulation studies for faults and cyberattacks scenarios, and finally in Sec. 5 we present the concluding remarks and scope for future work.

Notations: The following notations have been used in this work: \( I_n \) is an identity matrix of size \( n \), \( R(M) \) represents the range of matrix \( M \), \( B^\top \) represents the generalized inverse of matrix \( B \), \( \| \eta \| \) represents the Euclidean norm of the vector \( \eta \), and \( \| P \|_F \) represents the Frobenius norm of a matrix \( P \).

2 Problem Setup

2.1 Cyber-Physical System Model. Cyber-physical systems comprise of six layers: physical layer, control layer, communication layer, network layer, supervisory layer, and management layer [17]. The physical layer here represents the physical plant, sensors, and actuators. The network and communication layers contain the ICT and provides interconnections between the physical layer, control layer, and supervisory-management layers. Now, the control layer consists of the control module (CM) which contains controllers and state estimators. In contrast, the supervisory layer and management layer consist of the central management system managed through failure mode and effect analysis and strategies for management in terms of operating condition commands to the control layer, and supervisory-management layers. Now, the control layer, and supervisory-management layers. Now, the

Physical plant. Subsequently, let us consider the following linear time-invariant state space model for the physical plant \( S \):

\[
S : \quad \dot{x} = Ax + Bu + \eta, \quad y = x
\]

where \( x \in \mathbb{R}^n \) represents the states of the system; \( u \in \mathbb{R}^p \) represents the control input obtained from CM; \( A \in \mathbb{R}^{n \times n} \) represents the state matrix; \( B \in \mathbb{R}^{n \times p} \) is the actuation distribution matrix; \( \eta \in \mathbb{R}^p \) represents any unknown input. In this work, we have assumed full state feedback and thus the measurement \( y = x \). Our goal in this work is to focus on isolating cyberattack and faults on system actuation only, since it directly affects the system states. This isolation is crucial as cyberattacks on actuation cannot be bypassed, while corrupted sensors can be avoided by implementing sensor redundancies.

Notably, under a cyber-physical setting, the adversary is able to manipulate the actuation channel of the system to launch a cyberattack \( \alpha \in \mathbb{R}^p \). Consequently, such cyberattack on the actuation channel is modeled in the control-theoretic framework as an unknown input \( \eta = Bu \), since the control command passes through the actuation distribution matrix [17]. On the other hand, if there is a fault \( f \in \mathbb{R}^p \) in the system, then the unknown input \( \eta = Ef \), where \( E \in \mathbb{R}^{p \times p} \) represents the fault distribution matrix. It is to be noted that such distribution matrix \( E \) can be reliably obtained using failure mode and effect analysis and strategies for uncertainty quantification [18]. We present the schematic of our problem framework in Fig. 1.

Assumption 1. In this framework, we assume that these unknown inputs are injected to the system either as cyberattacks or faults but never both simultaneously, since model-based detection-isolation schemes are unable to handle such scenarios due to the inherent ill-posedness of estimating unknown inputs to the system dynamics or measurement. Even though some work on multi-agent systems have considered simultaneous cyberattacks and faults (largely limited to the connection topology), for general CPS systems only transient faults and cyberattacks have been considered simultaneously [9]. Moreover, this assumption ensures that the injected cyberattack is not covert, which can evade detection [4]. We also assume here that the unknown input is bounded such that

\[
\| \eta(t) \| < M < \infty, \quad \forall t
\]

In the next section, we propose a sliding mode-based diagnostic filter that detects and estimates unknown inputs defined in Eqn. (1).

2.2 Sliding Mode-Based Diagnostic Filter. In this formulation, the objective of the sliding mode-based DF is to detect and estimate unknown inputs (such as cyberattacks and faults) to the system. The structure of the filter considered here is based on measurement feedback from the system [19,20] and the unknown input \( \eta \) is estimated using an equivalent output error injection term. Let us first present the structure of the sliding mode-based DF as

\[
\bar{D}F: \quad \dot{x} = Ax + Bu + L(y - \hat{y}) / \| y - \hat{y} \|, \quad \dot{\hat{y}} = \hat{\gamma}
\]

where \( L \in \mathbb{R}^{p \times p} \) is the filter gain. Next, let us now define the error state as \( e = x - \hat{x} \). The error dynamics is then

\[
\dot{e} = Ae + \eta - L(x - \hat{x}) / \| x - \hat{x} \|
\]

For \( 0 < \tau < 1 \), the switching hyperplane and reaching law for the sliding mode-based DF (3) is respectively as

\[
S = \{ e \in \mathbb{R}^p : e = 0 \}, \quad \tau S = -e + (\tau - 1) \text{sgn}(e)
\]

Proposition 1 (Convergence of the sliding mode-based diagnostic filter). Consider the system given by Eq. (1) and the sliding mode-based DF given by Eq. (3). If there exists positive definite matrices \( P \in \mathbb{R}^{n \times n} \) and \( Q \in \mathbb{R}^{p \times p} \), constant \( \gamma > 0 \) such that

\[
A^T P + PA \leq -Q, \quad \text{and} \quad \gamma > \| P \|_F M
\]

where \( M \) is obtained from Eq. (2) and we choose filter gain \( L \) such that \( L = \gamma P^{-1} \), then the estimate for the bounded unknown input vector \( \eta \) is given by

\[
\hat{\eta} = F(L(x - \hat{x}) / \| x - \hat{x} \|)
\]

where \( F(.) \) is a low-pass filter function and \( \gamma \rightarrow \hat{\eta} \) in finite time. Proof. Let us define a Lyapunov functional \( V(e) = e^T Pe \), where \( P \) is determined by Eq. (6). Taking the time derivative of \( V(e) \) and using Eq. (4), we obtain

\[
\dot{V} = e^T (A^T P + PA)e + 2e^T P\eta - 2e^T PLe / \| e \|)
\]

With filter gain \( L = \gamma P^{-1} \) and using Eq. (6), we obtain

\[
\dot{V} \leq -\gamma e^T Qe + 2\| e \| \| P \|_F M - 2\gamma e^T Q e
\]
Since $-e^t Qe < 0$ due to positive definiteness of $Q$, we can write Eq. (8) as $V \leq 2[\|P\|_2 M + 2\|e\|]$. Considering the fact that $\gamma > \|P\|_2 M$, we can write $V < -2\beta_1 \|e\|$ where $\beta_1 = (-\|P\|_2 M + \gamma) > 0$. From this, and considering $V \geq \lambda_{\max}(P)\|e\|^2 \leq \sqrt{V}/\sqrt{\lambda_{\max}(P)}$, we can further write $V < -\beta_1 \sqrt{V}$ where $\beta_2 = 2\beta_1 / \sqrt{\lambda_{\max}(P)} > 0$. This implies that $e \to 0$ as $t \to T_{\text{max}}$ where $T_{\text{max}} < \infty$ is a finite time [20]. Consequently, after $t > T_{\text{max}}$, Eq. (4) becomes $0 = \eta - L(x - \hat{x})/\|x - \hat{x}\|$, and yields the equivalent output error dynamics [19,20]. Subsequently, we can use this solution to obtain an estimate of the unknown input by passing the output error injection term through a low-pass filter $\tilde{F}(\cdot)$ [19,20]. To achieve the reaching law, the low-pass filter is chosen as $\hat{\eta} = -(1/\eta_1 + (1/\tau)(L(y - \bar{y})/\|y - \bar{y}\|))$.

3 Distinguishability

Using the sliding mode-based DF from Proposition 1, we use the estimation of the unknown input $\eta(t)$ for distinguishing between cyberattack and fault. Now, if $\eta$ is a cyberattack, then $\eta = Ba$, i.e., $R(B)$ represents the plausible set of cyberattacks. Similarly, $R(E)$ represents the plausible set of faults. Therefore, the question of distinguishability translates to identifying if $\eta$ lies in $R(B)$ or $R(E)$. With this intent, we define the following Distinguishability metric.

DEFINITION 1 (Distinguishability metric). For an unknown input $\eta \neq 0$, we define a distinguishability metric

$$\mathcal{M}(\eta) := \|P_B\eta\|_2 - \|P_E\eta\|_2$$

(9)

In the above definition, $P_B := (I - B(B^T B)^{-1}B^T)$, $P_E := (I - E(E^T E)^{-1}E^T)$, $\|P_B\eta\|_2$ denotes the minimum distance of $\eta(t)$ from $R(B)$, while $\|P_E\eta\|_2$ denotes the minimum distance of $\eta(t)$ from $R(E)$. Hence, $\mathcal{M}$ represents how closer or further $\eta$ is to the range space of $B$ or $E$. Evidently, positive $M$ implies that the distance of $\eta(t)$ to $R(B)$ is more than the distance of $\eta(t)$ to $R(E)$. Similarly, if $\eta$ is negative, it implies $\eta$ is closer to the $R(B)$. Thus, $\mathcal{M} = 0$ for a non-zero $\eta$ and the unknown input is equidistant from both $R(B)$ and $R(E)$.

Remark 1. For an arbitrarily accurate estimation of the distribution matrices $B$, $E$, and the unknown input $\eta$, we will have only two scenarios. Either there exists an $\alpha$ such that $\eta = Ba \equiv 0$, or an $f$ such that $\eta - Ef \equiv 0$. This would have unambiguously proven that the unknown input $\eta$ is in fact a cyberattack in the first case and a fault for the second. However, such assumptions are unrealistic for detector performance and due to uncertainties in system model and measurements. Hence, the metric Eq. (9) is defined to obtain the degree of closeness of the unknown input $\eta$ to the space of plausible cyberattacks $R(B)$ and faults $R(E)$. Figure 2 shows the geometric interpretation of the distinguishability metric $\mathcal{M}$.

THEOREM 1 (Distinguishability criterion). Consider the system Eq. (1) with non-zero unknown input $\eta(t)$ and the sliding mode-based DF from Eq. (3). Let us also assume that the DF sati es conditions provided in Proposition 1. Then this estimated unknown input Eq. (7) is a cyberattack if $M(\hat{\eta}) < 0$ and is a fault if $M(\hat{\eta}) > 0$. Inversely, if $M(\hat{\eta}) = 0$, unknown inputs are indistinguishable as a fault or cyberattack.

Proof. The proof follows from Eq. (9).

COROLLARY 1 (Fault-mimicking cyberattacks). A cyberattack will be indistinguishable from a fault if there exists an $\alpha_* \in \mathbb{R}^n$ such that $\alpha_* = B^T Ef$, for some $f \in \mathbb{R}^r$, and such attacks are called fault-mimicking cyberattacks.

Remark 2. Constrained by the assumption of an uncertain CPS, Theorem 1 provides only a sufficient condition for distinguishability of attacks and faults. Hence, in Corollary 1, we present how the adversary can craft a fault-mimicking or fault-based cyberattack by leveraging this limited system knowledge of the defender. Generally, such attacks are often encountered in power grid or IoT-enabled CPS and Corollary 1 rectifies the fundamental limitation of the proposed criteria. To further address the concerns for uncertain CPS, our next theorem introduces a practical guarantee for distinguishability.

THEOREM 2 (Guaranteed practical distinguishability). Consider a system with bounded uncertainty where $\hat{\eta} \to \eta + e$. If the error in estimation $\|e\|_2$ is bounded by $M_\varepsilon$, then the estimated unknown input $\hat{\eta}$ is a cyberattack if $M(\hat{\eta}) \leq -\|P_E\|_2 M_\varepsilon$ and is a fault if $M(\hat{\eta}) > \|P_B\|_2 M_\varepsilon$.

Proof. Let us consider the case of cyberattack for uncertain CPS, where Eq. (9) yields $M(\hat{\eta}) - \|P_B\|_2 \geq -\|P_E\|_2 (\eta + e) \|_2 < 0$. Similarly for faulty scenario, we can show $M(\hat{\eta}) + \|P_E\|_2 > 0$. Next, we apply the bound of bounded uncertainty on $e$ and remove the overlap between these inequalities to obtain the condition of guaranteed distinguishability.

4 Simulation Results

In this section, we illustrate the proposed concepts using simulation studies. The system considered is given as: $A = -10B^T I_2$, $B = [3 \ 2]^T$, $D = 0$, and $E = [2 \ 5]^T$. The unknown input is $\eta = [\eta_1, \eta_2]^T$. For all the case studies in this section, a dynamic input profile is given as input $u$ to the system Eq. (1) (shown in Fig. 3). The observer gain for the sliding mode-based DF Eq. (3) is chosen to be $L = 50I_2$ and the filter gain as $r = 0.1$. In this simulation study, we consider two cases. For case 1, the system is under a physical fault. While for case 2, the system is subjected to a cyberattack. The objective of this case study is to show how the unknown input estimated by the sliding mode-based DF can be successfully identified as either fault or cyberattack using the distinguishability metric and criterion proposed in Eq. (9) and Theorem 1. With this, let us look at the results of the two case studies.

4.1 Case 1: Fault. For this case, the system is subject to a fault of magnitude $f = 5(1 - \exp(-10^{-2} t))$ and it is manifested to the system as an unknown input $\eta = Ef$. Figure 4 shows that the sliding mode-based DF can estimate this unknown input while starting from arbitrary initial conditions (shown in inset). The estimates of the two components of $\eta = [\eta_1, \eta_2]^T$ match with the true components of the unknown input $\eta = [\eta_1, \eta_2]^T$ (shown in Fig. 4).

Fig. 2 Realization of distinguishability metric $\mathcal{M}$ using the distance from (a) fault space and (b) attack space

Fig. 3 Dynamic input signal $u$ to the system Eq. (1)
The DF also estimates the states of system $\hat{x}_1$ and $\hat{x}_2$, and these estimates match significantly with the true states $x_1$ and $x_2$. We observe in Fig. 5 that $M > 0$ in steady-state, indicating that the unknown input is a fault. We also note here that the distinguishability metric is non-positive for the first 0.2 s (as shown in the inset of Fig. 5). However, this is due to the time needed by the DF in order to converge to the correct estimates of the states and unknown inputs to the system.

### 4.2 Case 2: Cyberattack

For this case study, we construct a cyberattack $\alpha$ that can drive the system states to some unintended states $[1, 1]$. Using steady-state condition of the system equation (Eq. (1)), we obtain the attack policy to be $\alpha = -B^T \hat{A} [1, 1]^T - \mu$. From Fig. 6, it is evident that the sliding mode-based DF can faithfully estimate the two components of the unknown input $\eta = [\eta_1, \eta_2]^T = B \alpha$. The initial condition for the unknown inputs are unspecified. Hence, the estimates are initialized arbitrarily. However, the DF converges to correct estimates in the steady-state starting from the arbitrary initial conditions (as shown in the inset of Fig. 6). Subsequently, we calculate the distinguishability metric $M$ from Eq. (9) and plot in Fig. 7. Since $M < 0$, we can conclude from the distinguishability criterion that the unknown input $\eta$ is a cyberattack.

### 5 Conclusion and Future Work

Distinguishing between the occurrence of faults and cyberattacks in a system is of utmost importance in order to provide appropriate mitigation strategies. In this work, we have proposed a mathematical framework for distinguishing between the two, utilizing a sliding mode-based DF. Using the estimate of the unknown input from DF, we proposed a distinguishability metric and criterion in order to achieve our goal. Finally, we have conducted a set of simulation studies with both faults and cyberattacks to illustrate the validity of our proposed framework. In the future, we intend to extend this work to nonlinear systems and consider different attack types such as denial-of-service, replay attacks, etc., through both control command and sensor measurement.

### Conflict of Interest

There are no conflicts of interest.

### Data Availability Statement

The datasets generated and supporting the findings of this article are obtainable from the corresponding author upon reasonable request.

### References


