2024 European Control Conference (ECC)
June 25-28, 2024. Stockholm, Sweden

Dynamic Adaptation Gain for Threat Discrimination

Kangkang Zhang, Kaiwen Chen, Marios M. Polycarpou, and Thomas Parisini

Abstract— **This paper proposes an adaptive-observer-based threat discrimination method for systems with disturbances, aiming to identify the occurring threat type: component faults or cyber attacks. Stealthy attacks are typically exponentially decaying with time and only slightly alter the system outputs, the effects of which can be easily inundated by non-attacker-incurred disturbances. To this end, an integrator is applied to the system output to retain the effects of stealthy attacks. Compared to the classical integral-type adaptive observers with a constant adaptation gain, a dynamic adaptation gain is exploited to provide additional degrees of design freedom for frequency-domain loop-shaping. This allows to apply distinct threat/disturbance-to-residual gains in the frequency intervals to which the threats and the disturbances belong, respectively, thereby improving the threat discrimination performance. A numerical example to demonstrate the effectiveness of the proposed methodology is presented.**

I. Introduction

Cyber-physical systems (CPS) are composed of complex interconnections between physical and cyber components, which are vulnerable to various threats, ranging from traditional physical faults to cyber attacks being studied in the more recent literature (see, *e.g.*, [1]). Techniques to determine the types of occurring threats (*i.e.,* threat discrimination) have become an urgent need of industrial practice and a challenge in engineering research.

Consider a threat that has been detected by some advanced anomaly detection methodologies such as fault diagnosis methodologies in [2] and the attack detection methodologies in [3], [4]. The threat discrimination algorithms are designed to determine the occurring threats (cyber attacks or physical faults). Typical fault isolation schemes that locate faulty components are in general not suitable for threat discrimination purposes since physical faults and cyber attacks may occur individually or simultaneously on the same component and a pure isolation result cannot provide sufficient information to reveal the nature of the occurring threats.

To achieve threat discrimination between physical faults and cyber attacks, an active approach based on the watermark technique is proposed in [5]. The injected watermark, however, degrades the control performance, which is the downside of most active approaches. Passive approaches preserve the original control performance, nevertheless, are more difficult to achieve good discrimination performance since only the measurement from the system can be exploited. A major challenge in passive threat discrimination is caused by attacks with vanishing amplitude, for example, stealthy attacks [6], [7] that exponentially decay with time and only slightly alter the system outputs, the effects of which can be easily inundated by disturbances and noise. How to improve the effectiveness of passive approaches for threat detection and discrimination, particularly in the presence of stealthy attacks, has been of increasing interest recently. By using passive adaptive observer techniques, [8] proposes a discrimination methodology that can identify faults from the potential fault scenarios though it cannot identify stealthy attacks. The scheme in [9] is also based on an adaptive observer and is capable of identifying both stealthy attacks and constant faults. Passive discrimination of stealthy attacks and time-varying physical faults remains an open problem, to the best of the authors' knowledge, and is to be studied by this paper.

In this paper, the threat discrimination problem for systems with disturbances is addressed by considering threats including stealthy cyber attacks and general physical faults. An adaptive observer-based threat discrimination strategy that is able to handle stealthy integrity attacks, is proposed. Dual adaptive observers serving as threat discriminators with dynamic adaptation gains (the concept introduced in [10]) are exploited to provide additional design freedoms in structure. Compared with integral-type adaptive observers using constant gains, the proposed structure of adaptive observer provides additional flexibility in frequency loop shaping and therefore improves threat discrimination capabilities. A loop-shaping method to tune the dynamic adaptation gain has been presented, which allows the balance of high sensitivity to threats and robustness to disturbances.

Notations: For a constant vector $x \in \mathbb{R}^n$, $\|x\|_2^2 = x^T x$. For a time-varying signal $x(t) \in \mathbb{R}^n$ and a finite time interval $[\tau, \tau + T_0]$, the root mean square is defined as $\|x(t)\|_{\text{RMS}}^2 = \frac{1}{T_0}\int_\tau^{\tau+T_0} x^T(t)x(t)\mathrm{d}t$. For a constant matrix $A \in \mathbb{R}^{n\times m}$, $\|A\|_2 = \sigma_{\max}(A)$ where $\sigma_{\max}$ represents the maximum singular value.

II. Problem Formulation

A. Cyber-Physical System Description

Two types of threats: cyber attacks and physical faults, are considered in this paper, and their combination results

This work has been supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101027980 (CSP-CPS-A-ICA), and No. 739551 (KIOS CoE-TEAMING), and by the EPSRC grant EP/X033546/1.

K. Zhang and K. Chen are with the Dept. of Electrical and Electronic Engineering, Imperial College London, London, SW7 2AZ, U.K. (e-mail: {kzhang5, kaiwen.chen16}@imperial.ac.uk

T. Parisini is with the Dept. of Electrical and Electronic Engineering, Imperial College London, London, SW7 2AZ, U.K., with the Dept. of Engineering and Architecture, University of Trieste, Trieste, 34127, Italy, and with the KIOS Research and Innovation Center of Excellence, Cyprus, (e-mail: t.parisini@gmail.com).

M. Polycarpou is with the KIOS Research and Innovation Center of Excellence and the Dept. of Electrical and Computer Engineering, University of Cyprus, Nicosia, 1678, Cyprus, (e-mail: mpolycar@ucy.ac.cy).

in three threat scenarios: the attack-only scenario, the fault-only scenario and the fault-and-attack scenario. For concise presentation, we denote by t_0 the threat occurring time, and let $\beta(t-t_0):\mathbb{R}_{\geq 0}\to\{0,1\}$ be the indicator function, *i.e.*, $\beta(t-t_0)=1$ for $t\geq t_0$, and otherwise, 0. Two indicators: $\beta^a\triangleq\beta^a(t-t_0)$ and $\beta^f\triangleq\beta^f(t-t_0)$, are defined to indicate the occurrence of attacks and faults respectively. Then, the pair $\beta\triangleq(\beta^f,\beta^a)$ indicates the threat scenario, and satisfies

$$\beta\in\{(1,0),(0,1),(1,1)\}. \tag{1}$$

Note that $\beta=(0,0)$ is not considered since the threat discrimination process is initiated only after a threat is detected.

Consider the dynamical system under the aforementioned threats described by

$$\dot{\chi}=A\chi+B\mu+\beta^aB^aa(t)+\beta^f(f_p(t)+Bf_a(t))+B_d\eta(t), \tag{2a}$$
$$\zeta=C\chi+D\mu+\beta^a\bar{D}^aa(t)+\beta^ff_s(t)+D_d\eta(t), \tag{2b}$$

where $\chi\in\mathbb{R}^{n_x}$ is the state of the physical plant with $\chi(t_0)=\chi_0$; $\mu\in\mathbb{R}^{n_u}$ is the control input; $\zeta\in\mathbb{R}^{n_y}$ represents the sensor measurement received from the network. A, B, C, B_d and D_d are known by the defender, and the pair (A,C) is observable. The $\eta\in\mathbb{R}^{n_d}$ represents the lumped disturbances and measurement noise, satisfying the following assumption.

Assumption 1. Let $d(t)\triangleq\int_0^t\eta(t)\mathrm{d}t$. Then, d is assumed to be uniformly bounded. Moreover, for a given time length T_0, there exists a constant $\delta_d>0$ such that

$$\frac{1}{T_0}\int_\tau^{\tau+T_0}d^T(t)d(t)\mathrm{d}t\leq\delta_d,\ \forall\ \tau\geq 0, \tag{3}$$

where δ_d is known by the defender. ▲

Note that Assumption 1 characterizes an average energy property of $d(t)$. A variety of signals satisfies Assumption 1 such as harmonic oscillations and white noise.

In the fault scenario, multiple types of faults are taken into consideration. In (2), f_p, f_a and f_s represent the process, actuator, and sensor faults, respectively. Let $f(t)\triangleq[f_p^T(t),f_a^T(t),f_s^T(t)]^T\in\mathbb{R}^{n_f}$, which allows re-writing $f_p(t)+Bf_a(t)=B^ff(t)$ in (2a) with $B^f=[I,B,0]$, and $f_s=D^ff$ in (2b) with $D^f=[0,0,I]$. In addition, f satisfies the following assumption.

Assumption 2. There exists a set of known and bounded regressors $\varphi_{f,1}(t),\cdots,\varphi_{f,n_f}(t)\in\mathbb{R}$ such that

$$f(t)=\varphi^f(t)\theta^f, \tag{4}$$

where $\varphi^f(t)=\mathrm{diag}\left(\varphi_{f,1}(t),\cdots,\varphi_{f,n_f}(t)\right)\in\mathbb{R}^{n_f\times n_f}$ and $\theta^f\in\mathbb{R}^{n_f}$ is the unknown parameter vector. ▲

Assumption 2 essentially requires that the potential fault can be parametrized by some known time-varying regressors, which can be determined through experiments. In practice, this parametrization does not need to be perfect since the effects of the parametrization error can be modelled by process disturbances.

B. Attack Scenarios

The signal $a(t)\in\mathbb{R}^{n_a}$ in (2) represents the cyber attacks. More specifically, a type of stealthy integrity attacks (referred to as *undetectable attacks* in [6]) are considered and characterized by the following equation

$$\zeta(t,t_0,\chi_0,\beta^ff,\eta,a)=\zeta(t,t_0,\chi_0+\Delta\chi_0,\beta^ff,\eta,0),$$
$$\forall\ t\geq t_0,\ f\in\mathbb{R}^{n_f},\ \eta\in\mathbb{R}^{n_d},\beta^f\in\{0,1\}, \tag{5}$$

for some $\Delta\chi_0\neq 0$. The left and right-hand sides represent the output trajectories of the system (2) driven by the initial condition χ_0, β^ff, η and $\beta^a=1$, and by the initial condition $\chi_0+\Delta\chi_0$, β^ff, η and $\beta^a=0$, respectively. Some specific attacks, for example, the zero dynamics attacks in [7], the covert attacks in [6] and the replay attacks in [11], [12] satisfy (5). We only consider the case in which $\Delta\chi_0\neq 0$. For the case in which $\Delta\chi_0=0$, the attacker is required to be able to disrupt all sensor and actuator communication channels, which is impractical and therefore not considered in this paper (interested readers can refer to [11] for details).

C. Problem Setup

Recalling (5), the effect of a considered attack on the system output can be described by an impulsive response. Instead of studying the system (2) under the attack, an integrator is introduced to compute the integral of the output signal ζ and retain the effect of the stealthy attacks. Then, by connecting the output of (5) to an integrator, the new output y is generated by the following dynamical system:

$$\dot{x}=Ax+Bu+\beta^aB^a\phi^a\theta^a+\beta^fB^f\phi^f\theta^f+B_dd, \tag{6a}$$
$$y=Cx+Du+\beta^aD^a\phi^a\theta^a+\beta^fD^f\phi^f\theta^f+D_dd. \tag{6b}$$

where $B_a=I$ and $D_a=0$, and x is a new state $x\neq\chi$ or $x\neq\int_{t_d}^t\chi d\tau$. In the system (6), y and u, ϕ^a and ϕ^f are available signals for discriminator design. It is worth pointing out that the difference of the frequency spectrums of φ^f and φ^a is retained in ϕ^f and ϕ^a, respectively. Moreover, $\phi^a(t)$ does not converge to zero and is unvanished as expected.

This paper aims to develop a threat discriminator such that when initiated at t_d, it can identify online the occurring threat scenario from the prescribed threat scenario set (1). The detailed objectives are summarized below:

(i) Design two specific adaptive observer-based discriminators that are equipped with a so-called dynamic adaptation gain, allowing additional degrees of freedom in structure for tuning the sensitivities to threats and robustness to disturbances;

(ii) Tune the dynamic adaptation gain and perform loop-shaping for the residual sensitivity so as to manifest the threats and suppress the disturbances and noise, according to their distinct spectrums.

III. Threat Discrimination Design

A. Threat Discrimination Strategy

The proposed threat discrimination scheme is composed of an attack discriminator $\mathbf{D}^a$ and a fault discriminator $\mathbf{D}^f$. Each discriminator $\mathbf{D}^s$, $s\in\{a,f\}$, is equipped with an

adaptive observer $\mathbf{O}^s$. The adaptive observer $\mathbf{O}^s$ uses ϕ^s as the regressor of its estimation model. Once the discriminators are activated, the adaptive observer $\mathbf{O}^s$, $s \in \{a, f\}$ adopts its estimation model with the associated regressor ϕ^s to estimate the threat function. Ideally, when the regressor ϕ^s matches the threat regressor, the generated residual r^s goes below the threshold J^s_{th}, while the residual r^s goes above J^s_{th} when the ϕ^s does not match the threat regressor. Consequently, the discrimination problem boils down to observing which residual signal triggers the mismatch condition.

B. Adaptive Discriminator Design

The threat discriminators are activated once any threat is detected at t_d. Let $s \in \{a, f\}$ denote the threat type. Motivated by the adaptive law with a dynamic adaptation gain in [10], We start by designing the adaptive observer for the system (6) with the threat s as follows:

$$\dot{\hat{x}}^s = A\hat{x}^s + Bu + B^s\phi^s\hat{\theta}^s + L^s\epsilon^s, \tag{7a}$$

$$\dot{\hat{p}}^s = \text{Proj}\left(\hat{p}^s,\ A^s_p\hat{p}^s + B^s_p\phi^s(t)\epsilon^s, \Gamma^s\right), \tag{7b}$$

$$\hat{\theta}^s = C^s_p\hat{p}^s, \tag{7c}$$

$$\hat{y}^s = C\hat{x}^s + D^s\phi^s(t)\hat{\theta}^s, \tag{7d}$$

where $\hat{x}^s \in \mathbb{R}^{n_x}$ and $\hat{y}^s \in \mathbb{R}^{n_y}$ are the estimates of x and y of the system (6), respectively, and $\hat{\theta}^s \in \mathbb{R}^{n_s}$ is the estimate of $\beta^s\theta^s$, $\epsilon^s \triangleq y - \hat{y}^s$ denotes the output estimation error. The matrix $C^s_p = [0, I_{n_s}]$, and $L^s \in \mathbb{R}^{n_x\times n_y}$, $A^s_p \in \mathbb{R}^{n_p\times n_p}, B^s_p \in \mathbb{R}^{n_p\times n_s}$, are design parameters. The matrix $\Gamma^s = \Gamma^{sT} > 0$ is a learning gain matrix. The observer is activated at $t = t_d$ and the initial values of $\hat{x}^s$ and $\hat{\theta}^s$ are respectively chosen as $\hat{x}^s(t_d) = 0$ and $\hat{\theta}^s(t_d) = 0$. The projection operator "Proj" restricts the trajectories of $\hat{p}^s$ to a predefined compact and convex set, which is closely related to the set Θ^s to which θ^s belongs. More specifically, in the special case where $\hat{\theta}^s$ is restricted in a hypersphere $\mathcal{P}^s$ with radius M^s, Θ^s is selected as hypersphere with radius M^s. The parameter adaptation process is described by (7b), with

$$\text{Proj}\left(\hat{p}^s, \tau^s, \Gamma^s\right) = \left(I - \Pi^s\Gamma^s\frac{\hat{p}^s\hat{p}^{sT}}{\hat{p}^{sT}\Gamma^s\hat{p}^s}\right)\tau^s, \tag{8}$$

where $\tau^s \triangleq A^s_p\hat{p}^s + B^s_p\phi^s\epsilon^s$ and

$$\Pi^s \triangleq \begin{cases} 0, & \text{if } |\hat{p}^s| < M^s \text{ or } |\hat{p}^s| = M^s \text{ and } \hat{p}^{sT}\tau^s \le 0, \\ 1, & \text{if } |\hat{p}^s| = M^s \text{ and } \hat{p}^{sT}\tau^s > 0. \end{cases} \tag{9}$$

The general PAA with DAG provides more design parameters, thereby more freedoms, allowing to guarantee the convergence of the estimation errors and simultaneously improve the robustness of the observer to disturbance d and the sensitivity to the attacks and faults. A key requirement for a valid PAA candidate for asymptotic regulation/observation[1] is that there should exist an equivalent system that can generate arbitrary output signals identical to the output of the PAA after applying a constant shift $\theta \in \Theta$, under the same input signal to the PAA. Consider now a general form of linear PAAs as follows.

$$\dot{\hat{p}}^s = A^s_p\hat{p}^s + B^s_p v^s,\ \hat{\theta}^s = C^s_p\hat{p}^s + D^s_p v^s, \tag{10}$$

where $\hat{p}^s(t) \in \mathbb{R}^{n_p}$, $v^s(t) \in \mathbb{R}^{n_s}$, $\hat{\theta}^s(t) \in \mathbb{R}^{n_s}$, and $n_p \ge n_s$. The requirement for (10) to be a PAA candidate can be stated as: for all $\theta^s \in \Theta^s$, there exists some $p^s \in \mathbb{R}^{n_p}$, such that the equations

$$\dot{\tilde{p}}^s = A^s_p\tilde{p}^s + B^s_p v^s,\ \tilde{\theta}^s = C^s_p\tilde{p}^s + D^s_p v^s, \tag{11}$$

holds, where $\tilde{p}^s \triangleq \hat{p}^s - p^s$, $\tilde{\theta}^s \triangleq \hat{\theta}^s - \theta^s$. Eqs. (11) describe the equivalent adaptation error system. We now provide sufficient conditions for (10) to be a PAA candidate below.

Proposition 1. *System* (10) *is a PAA candidate if* $\Theta^s \subseteq C^s_p \ker A^s_p$.

Proof. The statement $\Theta^s \subseteq C^s_p \ker A^s_p$ can be equivalently stated as for any $\theta^s \in \Theta^s$, there exists $p^s \in \ker A^s_p$ such that $C^s_p p^s = \theta^s$. Computing the adaptation error system from (10) yields

$$\dot{\tilde{p}}^s = A^s_p\tilde{p}^s + A^s_p p^s + B^s_p v^s = A^s_p\tilde{p}^s + B_p v^s, \tag{12a}$$

$$\tilde{\theta}^s = C^s_p\hat{p}^s - \theta^s + D^s_p v^s = C^s_p\tilde{p}^s + D^s_p v^s, \tag{12b}$$

which is equivalent to (11), and hence proves the claim. □

Theorem 1. *System* (10) *is a PAA candidate if 1) 0 is an eigenvalue of* A^s_p *with geometric multiplicity equal to* n_s*; and 2)* $\left(A^s_p, C^s_p\right)$ *is observable.*

Proof. Condition 1) implies that there exists an $n_p \times n_s$, full-column-rank matrix V such that $A^s_p V = 0$. Noting condition 2) and invoking the PBH observability lemma yields

$$\text{rank}\begin{bmatrix}\lambda I - A^s_p \\ C^s_p\end{bmatrix}\Bigg|_{\lambda=0} = n_p. \tag{13}$$

Note also that

$$\begin{bmatrix}-A^s_p \\ C^s_p\end{bmatrix}V = \begin{bmatrix}0 \\ W\end{bmatrix}, \tag{14}$$

one can immediately see that W is a full-rank $n_s \times n_s$ matrix such that $C^s_p V = W$. Hence $C^s_p \ker A^s_p = \mathbb{R}^{n_s} \supset \Theta^s$ and invoking Proposition 1 proves the theorem. □

C. Convergence of Error Systems

This subsection will establish the convergence of error signals when the occurring threat matches the regressor in the observer (7). Let $\tilde{x}^s \triangleq x - \hat{x}^s$, and $\tilde{p}^s \triangleq p^s - [0, \hat{\theta}^{sT}]^T$. Then, in the presence of the threat $s \in \{a, f\}$, the error system is described as follows:

$$\dot{\tilde{x}}^s = A^s_0\tilde{x}^s + (B^s - L^sD^s)\phi^s(t)\tilde{\theta}^s + (B_d - L^sD_d)d, \tag{15a}$$

$$\dot{\tilde{p}}^s = A^s_p\tilde{p}^s + B^s_p\phi^s(t)\epsilon^s - \Pi^s\Gamma^s\frac{\hat{p}^s\hat{p}^{sT}}{\hat{p}^{sT}\Gamma^s\hat{p}^s}\tau^s, \tag{15b}$$

$$\tilde{\theta}^s = C^s_p\tilde{p}^s, \tag{15c}$$

$$\epsilon^s = C\tilde{x}^s + D^s\phi^s(t)\tilde{\theta}^s + D_d d, \tag{15d}$$

[1] We call it a PAA candidate for conciseness hereafter.

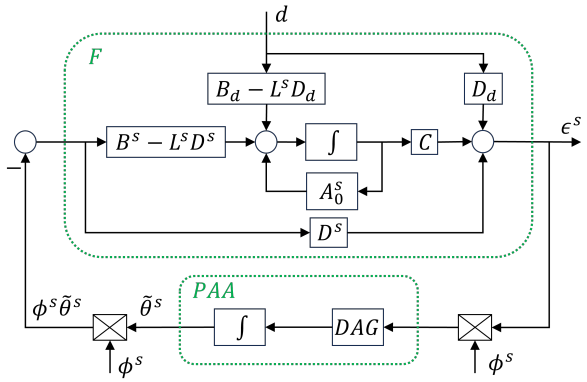

Fig. 1. Schematic representation of the closed-loop system.

where $A_0^s \triangleq A - L^s C$. An equivalent feedback representation of the error system (15) is depicted in Fig. 1. The boundedness of the signals in (15) is established by the theorem below.

Theorem 2. *In the threat $s \in \{a, f\}$ case, if there exist $L^s \in \mathbb{R}^{n_x \times n_y}$, $A_p^s \in \mathbb{R}^{n_z \times n_z}$, $B_p^s \in \mathbb{R}^{n_z \times n_i}$ such that*

(i) *(Strictly Positive Real Condition [13, Lemma 6.3]) there exists $P_0^s = P_0^{sT} > 0$, K_0^s and W_0^s, and a positive constant ε^s such that*

$$P_0^s A_0^{sT} + A_0^s P_0^s = -K_0^{sT} K_0^s - \varepsilon^s P_0^s, \tag{16a}$$
$$P_0^s (B^s - L^s D^s) = C^T - K_0^{sT} W_0^s, \tag{16b}$$
$$W_0^{sT} W_0^s = D^s + D^{sT}, \tag{16c}$$

(ii) *the matrices A_p and C_p satisfy the conditions in Theorem 1 and there exists $P_p^s = P_p^{sT} > 0$ and K_p^s such that*

$$P_p^s A_p^{sT} + A_p^s P_p^s = -K_p^{sT} K_p^s, \tag{17a}$$
$$P_p^s B_p^s = C_p^{sT}, \tag{17b}$$

then, in the absence of the disturbance, i.e., $d = 0$, the state estimation error $\tilde{x}^s$ converges to zero asymptotically. In the presence of the disturbance, the errors $\tilde{x}^s$, $\tilde{p}^s$, ϵ^s and $\tilde{\theta}^s$ are uniformly bounded.

The proof of this theorem is omitted due to the space limitation.

D. Loop Shaping Algorithm for Sensitivity and Robustness

In this subsection, a DAG tuning method is presented for shaping the loops of the attack and fault threat discriminators, thus, improving the sensitivity to threats and robustness to disturbances. Typically, the frequencies of threats have a narrow band in a low-frequency range, whereas the frequency spectrum of disturbances is usually distributed over a high-frequency range. To characterize the frequency spectrum of threats and disturbances, $\varpi_l > 0$ is used as the centre of the frequency band for the threats, and $[\varpi_h, \infty)$ with $\varpi_h > 0$ is used to represent the frequency range of the disturbances.

For two systems P_1 and P_2, we use $\mathcal{F}_l(P_1, P_2)$ to denote the closed-loop system with feedforward system P_1 and feedback system P_2. In addition, the residual for threat discrimination is chosen as the output estimation error of the adaptive observer, i.e., $r^s \triangleq \epsilon^s$. Let $\Delta\phi^{sh} \triangleq \phi^s - \phi^h$ where $s, h \in \{a, f\}$ and $s \neq h$, denote the difference between ϕ^s and ϕ^h. Then, in the presence of the threat case $h \in \{a, f\}$, r^s for $s \in \{a, f\}$ can be written as

$$r^s = \begin{bmatrix} F_1^s \mathcal{F}_l(1, \phi^s F_{PAA}^s \phi^s) & \mathcal{F}_l(F_2^s, \phi^s F_{PAA}^s \phi^s) \end{bmatrix} \begin{bmatrix} d \\ \Delta\phi^{sh}\theta^h \end{bmatrix}, \tag{18}$$

where $F_1^s \triangleq (A_0^s, B_d - L^s D_d, C, D_d)$ and $F_2^s \triangleq (A_0^s, B^s - L^s D^s, C, D^s)$. Moreover, $F_1^s \mathcal{F}_l(1, \phi^s F_{PAA}^s \phi^s)$ and $\mathcal{F}_l(F_2^s, \phi^s F_{PAA}^s \phi^s)$ represent the system from d and $\Delta\phi^{sh}\theta^h$ to r^s, respectively, which are depicted by (a) and (b) in Fig. 2, respectively.

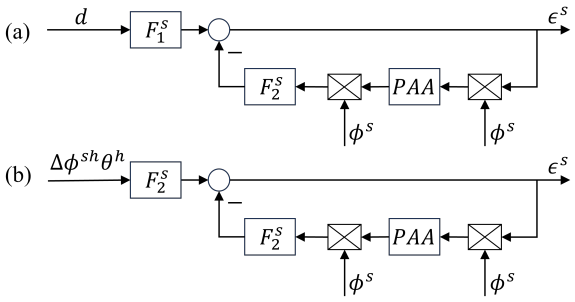

Fig. 2. Schematic representation of (a) system $F_1^s \mathcal{F}_l(1, \phi^s F_{PAA}^s \phi^s)$: disturbance to error path and (b) system $\mathcal{F}_l(F_2^s, \phi^s F_{PAA}^s \phi^s)$: mismatched regressor to error path.

For pre-determined F_1^s and F_2^s, the gains of $\mathcal{F}_l(1, F_2^s \phi^s F_{PAA}^s \phi^s)$ in the high and low-frequency ranges significantly affect the ability to discriminate the threats. To improve the discrimination abilities, we explore increasing the gain at ϖ_l and decreasing the gain in the frequency region $[\varpi, \infty)$. In the following, we focus on the single input and single output case of the system $\mathcal{F}_l(1, F_2^s \phi^s F_{PAA}^s \phi^s)$, and $\phi^s(t)$ is supposed to have an averaging constant 1. The superscript s is omitted in this section when no confusion is caused.

We proceed with the design in the frequency domain below, and start by writing F_2, F_{DAG} and F_{PAA} as transfer function form in the frequency domain as follows:

$$F_2 = \frac{N(j\omega)}{D(j\omega)}, \; F_{DAG} = \frac{R(j\omega)}{S(j\omega)}, \; F_{PAA} = \frac{F_{DAG}(j\omega)}{j\omega}, \tag{19}$$

where N, D, R and S are monic polynomials with respect to $j\omega$, $N_N, N_D, N_R, N_s \in \mathbb{N}_{\geq 0}$ are degrees of N, D, R and S, respectively. For given F_2, $N(j\omega)$ and $D(j\omega)$ are determined, $R(j\omega)$ and $S(j\omega)$ of the DAG are to be designed. For the system $\mathcal{F}_l(1, F_2 F_{PAA})$, the sensitivity function $P(j\omega)$ and the complementary function $M(j\omega) \triangleq 1 - P(j\omega)$ can be written as

$$P(j\omega) \triangleq \mathcal{F}_l(1, F_2 F_{PAA}) = \frac{j\omega D(j\omega) S(j\omega)}{W(j\omega)}, \tag{20a}$$
$$M(j\omega) \triangleq 1 - \mathcal{F}_l(1, F_2 F_{PAA}) = \frac{N(j\omega) R(j\omega)}{W(j\omega)}, \tag{20b}$$

where $W(j\omega)$ represents the polynomial specifying the desired closed-loop poles, and $W(j\omega)$, $D(j\omega)$ and $R(j\omega)$ satisfy the diophantine equation:

$$j\omega D(j\omega)S(j\omega) + N(j\omega)R(j\omega) = W(j\omega). \tag{21}$$

In this work, the desired closed-loop poles $W(j\omega)$ are divided into three parts:

$$W(j\omega) = W_D(j\omega)W_{H1}(j\omega)W_{H2}(j\omega), \tag{22}$$

where $W_D(j\omega)$ defines the dominant poles, $W_{H1}(j\omega)$ and $W_{H2}(j\omega)$ define the auxiliary poles. In order to improve the robustness to high-frequency disturbances, $W_{H1}(j\omega)$ are chosen as follows:

$$W_{H1}(j\omega) = (j\omega + \varpi_h)^{n_{W_{H1}}}, \tag{23}$$

where the degree $n_{W_{H1}} \in \mathbb{N}_{\geq 0}$. Furthermore, $R(j\omega)$ is divided into two parts as below:

$$R(j\omega) = R_D(j\omega)R_H(j\omega), \tag{24}$$

where $R_H(j\omega)$ is the pre-specified part of the DAG. To improve the sensitivity to threats at the specific frequency ϖ_0, the gain of the complementary sensitivity function $M(j\omega)$ must decrease at ϖ_0, which can be achieved by designing $R_H(j\omega)$ and $W_{H_2}(j\omega)$. In this work, we choose $R_H(j\omega)$ and $W_{H_2}(j\omega)$ to form a second-order notch filter [14], given as follows:

$$\frac{R_H(j\omega)}{W_{H2}(j\omega)} = \frac{(j\omega)^2 + 2j\xi_{num}\varpi_l\omega + \varpi_l^2}{(j\omega)^2 + 2j\xi_{den}\varpi_l\omega + \varpi_l^2}, \tag{25}$$

where $\xi_{num}, \xi_{den} \in (0, 1)$. An attenuation around ω_0 is obtained for $\xi_{num} < \xi_{den}$, and the maximal attenuation is given by

$$\min_{\omega\in\mathbb{R}_{\geq 0}} \left|\frac{R_H(j\omega)}{W_{H2}(j\omega)}\right| = \left|\frac{R_H(j\varpi_l)}{W_{H2}(j\varpi_l)}\right| = \frac{\xi_{num}}{\xi_{den}}. \tag{26}$$

Until now, the pre-specified terms R_H, W_{H1} and W_{H2} are determined. We turn to present the algorithm to determine R_D and S for the given W in Algorithm 1.

Algorithm 1: DAG design via loop shaping

Inputs : Desired pole polynomials $W_D(j\omega)$, Degree $n_{W_{H1}}$ of $W_{H1}(j\omega)$, damping ratios ξ_{num} and ξ_{den} $(0 < \xi_{num} < \xi_{den}) < 1$, center frequency of threats ϖ_l, and frequency range $[\varpi, \infty)$ of disturbances

Outputs: Zero polynomial $R(j\omega)$ of DAG and pole polynomial $S(j\omega)$ of DAG

1 Give zero and pole polynomials W_{H1}, W_{H2} and R_F:
$W_{H1}(j\omega) \leftarrow (j\omega - \varpi_j)^{n_{W_{H1}}}$,
$W_{H2}(j\omega) \leftarrow (j\omega)^2 + 2j\xi_{num}\varpi_l\omega + \varpi_l^2$,
$R_F(j\omega) \leftarrow (j\omega)^2 + 2j\xi_{den}\varpi_l\omega + \varpi_l^2$;

2 Get degrees of pre-specified zero and pole polynomials: $n_{W_D} \leftarrow \deg(W_D)$, $n_{R_F} \leftarrow \deg(R_F)$, $N_D \leftarrow n_{W_D} + n_{W_{H1}} + 2$;

3 Determine the degrees of Q-parameterization polynomial $Q(s)$: $n_{R_0} \leftarrow n_{R_F} + n_D$, $n_{R_0} \leftarrow n_D$, $n_Q \leftarrow n_{R_F} - 1$;

4 Solve the Diophantine equation: $R_D(j\omega)R_F(j\omega) - j\omega D(j\omega)Q(j\omega) = R_0(j\omega)$;

5 Compute $R(j\omega)$ and $S(j\omega)$:
$R(j\omega) = R_0(j\omega) + j\omega D(j\omega)Q(j\omega)$,
$j\omega D(j\omega)S(j\omega) + N(j\omega)R(j\omega) = W(j\omega)$.

IV. Discrimination Decision Rule

The threat discrimination decision rule is presented in this section. To this end, the residual, its evaluation and the corresponding adaptive threshold are determined. The output estimation error ϵ^s, $s \in \{a, f\}$ is chosen as the residual, i.e., $r^s \triangleq \epsilon^s = y - \hat{y}^s$, $\forall\ s \in \{a, f\}$. The root mean square of r^s in finite time interval T is chosen as the evaluation function, i.e., $J^s = \|r^s\|_{\text{RMS}}$. Then, an adaptive threshold can be chosen as

$$\begin{aligned} J_{th}^s(t) &\triangleq e^{-\lambda^s(t-t_d)}\delta_0 + \gamma^s\|d\|_{\text{RMS}}, \\ s.t.\ \gamma^s &= \max_{\omega\in[\varpi_h,\infty)} \|F_1^s(j\omega)P(j\omega)\|_2, \end{aligned} \tag{27}$$

where $\lambda^s > 0$ satisfies $|e^{A_0^s(t-t_d)}| \leq e^{-\lambda^s(t-t_d)}$, $\delta_0 \geq \|x(t_d)\|_2$, and P is given in (20a). From Assumption 1, we derive the adaptive threshold as follows:

$$J_{th}^s(t) = e^{-\lambda(t-t_d)}\delta_0 + \gamma^s\delta_d. \tag{28}$$

Based on the above defined r^s and J_{th}^s, we present the rule in Tab. I. The threat discrimination is done column-wise based on the second and third rows, and β given in (1) can then be determined and given in the fourth row.

TABLE I

Signature matrix of threat discrimination.

Residual & Threshold	Attack	Fault	Fault and Attack
$\|r^a(t)\|_{\text{RMS}} > J_{th}^a$	0	1	1
$\|r^f(t)\|_{\text{RMS}} > J_{th}^f$	1	0	1
β	(0,1)	(1,0)	(1,1)

V. Simulation Result

In this section, a numerical example is presented. The system matrices are given as follows:

$$A = \begin{bmatrix} -1 & 2 \\ 1 & 0 \end{bmatrix}, B = B^f = \begin{bmatrix} 1 \\ 0 \end{bmatrix},\ B_d = \begin{bmatrix} 1 \\ 1 \end{bmatrix},\ C = [\,1\ 0.5\,],\ D_d = 1.$$

In this simulation, the threat happens at $t_0 = 20$ s. The fault vector $f(t)$ satisfying Assumption 2 is assumed to be approximated by $\varphi(t) = \sin(2\pi \times 0.5t)$ with $\varpi_l = 2\pi \times 0.5$ rad/s. For the simulation purpose, the disturbance satisfying Assumption 1 is given by $d = \sin(2\pi \times 10t)$ with its frequency spectrum distributed in high frequencies and $\varpi_h = 2\pi \times 10$ rad/s.

To show the improvement of the threat discrimination ability by using the proposed adaptive observer with DAG, the singular values of the sensitivity function $P(j\omega)$ defined in (20a) for the adaptive observers with DAG and classical constant adaptation gain (Classical AG) are shown in Fig. 3, respectively. It shows that at the centre frequency of the fault,

i.e., $\varpi_l = 2\pi \times 0.5$ rad/s, the gain of the adaptive observer with DAG is much higher than the one with Classical AG. Their gains in the high-frequency range $[\varpi_h, \infty)$ are similar.

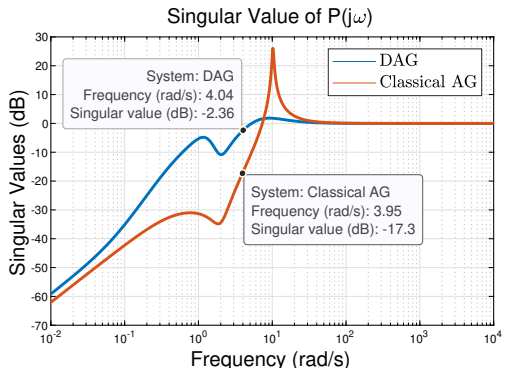

Fig. 3. Singular values of the sensitivity function $P(j\omega)$ for adaptive observers with DAG and Classical AG.

In this simulation, a stealthy zero-dynamics attack is used and is given by $a(t) = e^{-0.5(t-t_0)}$. The attack and/or the fault is considered to be detected by an anomaly detector at $t_d = 22$ s and thus, the discrimination schemes are activated at $t_d = 22$ s. The three threat scenarios considered in this paper are verified distinctively, and due to the space limitation, we present two of them in the following simulation results.

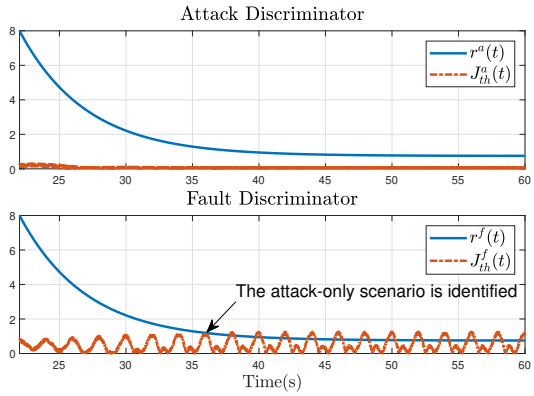

Fig. 4. Residuals and adaptive thresholds in the pure attack scenario.

1) Attack-only case: The discrimination result is shown in Fig. 4 and indicates by Tab. I that $\beta = (0, 1)$. Thus, we can conclude that the occurring threat type is an attack.

2) Fault-and-attack case: The discrimination result is shown in Fig. 5 and indicates based on Tab. I that $\beta = (1, 1)$. Hence, both fault and attack are occurring.

VI. Conclusion

In this paper, the threat discrimination problem for cyber-physical systems with disturbances has been studied. Threats including stealthy cyber attacks and general physical faults have been considered. An adaptive observer-based threat discrimination strategy that can handle stealthy integrity attacks, has been proposed. Dual adaptive observers serving as threat discriminators with dynamic adaptation gains

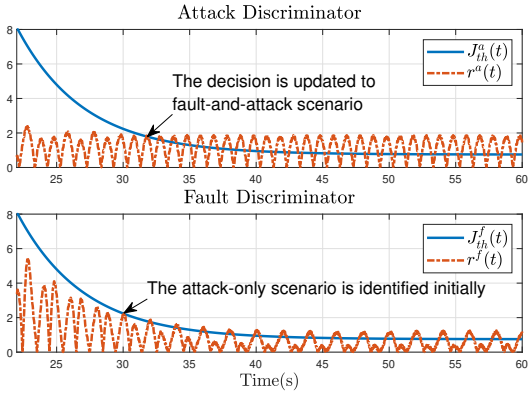

Fig. 5. Residuals and adaptive thresholds in the fault-and-attack case.

have been introduced to provide additional design freedoms in structure. Compared with adaptive observers with pure integral-type adaptation algorithms, the proposed structure of adaptive observer provides additional flexibility in frequency loop shaping and therefore improves threat discrimination capabilities. A loop-shaping method to tune the dynamic adaptation gain has been presented, which allows the balance of high sensitivity to threats and robustness to disturbances.

References

[1] A. Cardenas, S. Amin, and S. Sastry, "Secure control: Towards survivable cyber-physical systems," in *28th Int. Conf. Distrib. Comput. Syst. Workshops*. IEEE, 2008, pp. 495–500.

[2] S. X. Ding, *Model-based Fault Diagnosis Techniques: Design Schemes, Algorithms, and Tools*. Springer Science & Business Media, 2008.

[3] Y. Mo, R. Chabukswar, and B. Sinopoli, "Detecting integrity attacks on SCADA systems," *IEEE Trans. Control Syst. Technol.*, vol. 22, no. 4, pp. 1396–1407, 2013.

[4] P. Griffioen, S. Weerakkody, and B. Sinopoli, "A moving target defense for securing cyber-physical systems," *IEEE Trans. on Autom. Control*, vol. 66, no. 5, pp. 2016–2031, 2020.

[5] K. Zhang, A. Kasis, M. M. Polycarpou, and T. Parisini, "A sensor watermarking design for threat discrimination," *IFAC-PapersOnLine*, vol. 55, no. 6, pp. 433–438, 2022.

[6] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Trans. Autom. Control*, vol. 58, no. 11, pp. 2715–2729, 2013.

[7] A. Teixeira, I. Shames, H. Sandberg, and K. Johansson, "A secure control framework for resource-limited adversaries," *Automatica*, vol. 51, pp. 135–148, 2015.

[8] K. Zhang, C. Keliris, M. M. Polycarpou, and T. Parisini, "Discrimination between replay attacks and sensor faults for cyber-physical systems via event-triggered communication," *Eur. J. Control*, 2021.

[9] K. Zhang, C. Keliris, T. Parisini, and M. M. Polycarpou, "Identification of sensor replay attacks and physical faults for cyber-physical systems," *IEEE Control Syst. Lett.*, vol. 6, pp. 1178–1183, 2021.

[10] I. D. Landau, T.-B. Airimitoaie, B. Vau, and G. Buche, "On a general structure for adaptation/learning algorithms.—stability and performance issues," *Automatica*, vol. 156, p. 111193, 2023.

[11] K. Zhang, C. Keliris, T. Parisini, and M. M. Polycarpou, "Stealthy integrity attacks for a class of nonlinear cyber-physical systems," *IEEE Trans. Autom. Control*, vol. 67, no. 12, pp. 6723–6730, 2021.

[12] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *47th annu. Allerton Conf. Commun. Control, and Comput.* IEEE, 2009, pp. 911–918.

[13] H. Khailil, *Nonlinear Systems*. Prentice Hall, 2002.

[14] H. Procházka and I. D. Landau, "Pole placement with sensitivity function shaping using 2nd order digital notch filters," *Automatica*, vol. 39, no. 6, pp. 1103–1107, 2003.