

False Data-Injection Attack Detection in Cyber-Physical Systems: A Wasserstein Distributionally Robust Reachability Optimization Approach

Yulin Feng, Dapeng Lan, *Member, IEEE*, and Chao Shang, *Member, IEEE*

***Abstract*—Cyber-physical system (CPS) is the foundational backbone of modern critical infrastructures, so ensuring its security and resilience against cyber-attacks is of pivotal importance. This paper addresses the challenge of designing anomaly detectors for CPS under false-data injection (FDI) attacks and stochastic disturbances governed by unknown probability distribution. By using the Wasserstein ambiguity set, a prevalent data-driven tool in distributionally robust optimization (DRO), we first propose a new security metric to deal with the absence of disturbance distribution. This metric is designed by asymptotic reachability analysis of state deviations caused by stealthy FDI attacks and disturbance in a distributionally robust confidence set. We then formulate the detector design as a DRO problem that optimizes this security metric while controlling the false alarm rate robustly under a set of distributions. This yields a trade-off between robustness to disturbance and performance degradation under stealthy attacks. The resulting design problem turns out to be a challenging semi-infinite program due to the existence of distributionally robust chance constraints. We derive its exact albeit non-convex reformulation and develop an effective solution algorithm based on sequential optimization. Finally, a case study on a simulated three-tank is illustrated to demonstrate the efficiency of our design in robustifying against unknown disturbance distribution.**

***Index Terms*—Robust FDI attack detection, reachability analysis, performance degradation, distributionally robust optimization.**

I. Introduction

Cyber-Physical Systems (CPSs) are a critical component of modern technological advancements, integrating cyber communication and computation with physical plants. CPSs can be used in various fields, such as autonomous vehicles [1], industrial control systems [2] and power grid [3]. However, CPSs are often exposed to threats from external cyber-attackers, who may compromise communication networks, manipulate sensor data, or interfere with control signals, leading to severe consequences. Therefore, the security of CPS has gained increasing research attention over the recent decade.

As an essential problem of CPS security, anomaly detection aims at identifying unusual or malicious actions, thereby addressing the underlying challenge of mitigating malicious activities before significant harms are caused [4]. The anomaly detection problem is closely related to fault detection of technical processes, which has been a hot spot in the control community [5], [6]. However, cyber-attacks are more difficult to deal with than faults since attackers can meticulously design the attack mechanism to deceive the detector by exploiting available system knowledge [4], [7], [8]. Among various types of cyber-attacks, the stealthy false data-injection (FDI) attack has received the most attention and many tailored detector design schemes have been developed. In [7], a subspace-based detector based on coding theory was proposed to tackle undetectable sensor FDI attacks. [9] borrowed ideas from the random finite set theory to detect multiple attacks on different sensors. Deep reinforcement learning has also been found useful for detecting FDI attacks under disturbance by formulating the design problem as a partially observable Markov decision process in [10]. Aiming to accelerate the response speed, [11] designed an optimal weighting fusion criterion to calibrate the threshold under the limited bandwidth. A generalized likelihood ratio-based scheduler was presented in [12], which selectively transmits the most informative sensor data to detect potential cyber-attacks under limited communications.

Uncertain disturbance is widespread in real-world CPS. In previous works on cyber-attack detection, it is frequently assumed that the disturbance is either bounded or governed by a Gaussian distribution to ease analysis and design. Under the assumption of bounded disturbances, [13] proposed and minimized a valid reachability-based performance metric of the CPS, which is also useful for further guiding the control synthesis [14]–[16]. However, it critically hinges on the set-based description of system disturbances. In engineering practice, it remains non-trivial to obtain precise support information of stochastic disturbance. On the other side, the well-known generalized likelihood ratio test (GLRT) provides an optimal detection framework under the Gaussian assumption in [17]. However, once disturbance distribution deviates from Gaussianity, such as exhibiting heavy-tailed or non-stationary characteristics, the accuracy and reliability of cyber-attack detection are inevitably compromised. As an effective technique for managing uncertainty in probability distributions, distributionally robust optimization (DRO) has attracted extensive attention in the field of anomaly detection recently. Instead of assuming the true probability distribution to be precisely

This work was supported by the National Natural Science Foundation of China under Grant 62373211 *(Corresponding author: Chao Shang)*.

Yulin Feng and Chao Shang are with the Department of Automation, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: fyl23@mails.tsinghua.edu.cn; c-shang@tsinghua.edu.cn;).

Dapeng Lan is with the Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110003, China (e-mail: landapeng@sia.cn).

known, DRO optimizes the worst-case performance within the ambiguity set composed of all possible distributions, thereby offering a more robust solution than generic methods. While various distributionally robust anomaly detectors have already been developed, see e.g. [18]–[22], they are primarily oriented towards fault detection tasks, aiming to maximize sensitivity to faults but ignoring the adversarial nature of cyber-attacks.

This work aims at addressing the FDI attack detection problem in CPS subject to stochastic disturbance whose distribution is unknown. The main contributions of this work can be summarized as follows:

- We propose a new distributionally robust security metric to evaluate the performance degradation due to stealthy FDI attacks and stochastic disturbance governed by a unknown distribution. This is achieved by using the Wasserstein ambiguity set for data-driven uncertainty description and performing asymptotic reachability analysis.
- We formulate FDI attack detector design as a DRO problem that minimizes our proposed performance degradation metric while simultaneously controlling the false alarm rate (FAR), thereby achieving a desirable trade-off between security against attacks and robustness to disturbances.
- We establish an exact reformulation of the proposed distributionally robust detector design problem, successfully transforming a semi-infinite program into a finite-dimensional problem but involving bilinear matrix inequalities (BMIs). To solve this resulting non-convex problem, we develop a customized and efficient algorithm based on sequential optimization.

The remainder of this article is organized as follows. The basics of CPS and the distributionally robust anomaly detector design are revisited in Section II. Section III presents our attack detector design scheme, and case study results are given in Section IV. Section V concludes this article.

Notations: We use $\mathcal{N}(\mu, \Sigma)$ to denote a Gaussian distribution with mean μ and covariance Σ . χ^2_m denotes a Chi-square distribution with n degrees of freedom and its upper α-quantile is denoted by $\chi^2_m(\alpha)$. I denotes an identity matrix of appropriate dimension. For a vector x, the weighted l_2-norm $\|x\|_W = (x^\top W x)^{\frac{1}{2}}$ and $\|f\| = (f^\top f)^{\frac{1}{2}}$ with omitting the weight $W = I_{n_x}$. For a matrix X, $X_{[i:j]}$ denotes the submatrix X that goes from the ith to the jth column. $X^\dagger$ denotes its Moore–Penrose inverse, $\mathrm{tr}\{X\}$ denotes its trace, and $\|X\|_2$ denotes its spectral norm. $\mathrm{diag}\{X_1, \cdots, X_n\}$ is the block diagonal matrix with diagonal block matrices $X_1, \cdots, X_n$. For a symmetric X, $X \succeq (\succ) 0$ indicates that X is positive (semi-)definite. For discrete-time signal $x(k)$, the concatenated vector is denoted by $x_s(k) = \left[x(k-s+1)^\top \cdots x(k)^\top\right]^\top$.

II. Preliminaries

A. CPS Configuration

As described in Fig. 1, a CPS under study consists of a physical plant, a feedback controller, and an attack detector.

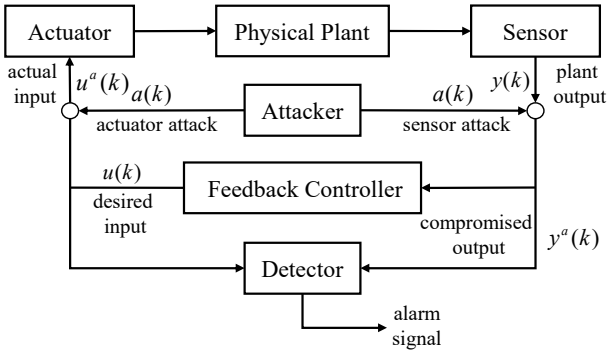

Fig. 1. CPS under FDI attacks.

The physical plant can be described as a stochastic discrete-time linear time-invariant (LTI) system:

$$\begin{cases} x(k+1) = Ax(k) + Bu^a(k) + B_d d(k) \\ y(k) = Cx(k) + D_d d(k) \end{cases} \tag{1}$$

where $x \in \mathbb{R}^{n_x}$, $y \in \mathbb{R}^{n_y}$, $u^a \in \mathbb{R}^{n_u}$ and $d \in \mathbb{R}^{n_d}$ denote system state, output signal, actual control signal, stochastic disturbance, and additive faults, respectively. A, B, B_d, C and D_d are state-space matrices of appropriate dimensions. It is assumed that (A, C) is observable and (A, B) is controllable. The sensor and actuator attacks can be described as:

$$\begin{cases} u^a(k) = u(k) + B_a a(k) \\ y^a(k) = y(k) + D_a a(k) \end{cases} \tag{2}$$

where the attack $a(k) \in \mathbb{R}^{n_u+n_y}$ signal injects false data into $u(k)$ and $y(k)$ through channels B_a and D_a, and $y^a(k)$ represents the compromised output. $u(k)$ is the desired input produced by the feedback controller:

$$\begin{cases} x_c(k+1) = A_c x_c(k) + B_c[y_{\mathrm{ref}}(k) - y^a(k)] \\ u(k) = C_c x_c(k) + D_c[y_{\mathrm{ref}}(k) - y^a(k)] \end{cases} \tag{3}$$

where $x_c \in \mathbb{R}^{n_c}$ and $y_{\mathrm{ref}}(k) \in \mathbb{R}^{n_y}$ are the state of the controller and the reference output signal. By defining the augmented state $\bar{x}(k) = \left[x(k)^\top \quad x_c(k)^\top\right]^\top$ and combining (1), (2) and (3), one attains the dynamics of the closed-loop system under attack:

$$\begin{cases} \bar{x}(k+1) = \bar{A}\bar{x}(k) + \bar{B}_{\mathrm{r}} y_{\mathrm{ref}}(k) + \bar{B}_a a(k) + \bar{B}_d d(k) \\ y^a(k) = \bar{C}\bar{x}(k) + D_a a(k) + D_d d(k) \end{cases} \tag{4}$$

where

$$\bar{A} = \begin{bmatrix} A - BD_cC & BC_c \\ -B_cC & A_c \end{bmatrix},\ \bar{B}_{\mathrm{r}} = \begin{bmatrix} BD_c \\ B_c \end{bmatrix},\ \bar{C} = \begin{bmatrix} C \\ 0 \end{bmatrix},$$
$$\bar{B}_a = \begin{bmatrix} B(B_a - D_cD_a) \\ -B_cD_a \end{bmatrix},\ \bar{B}_d = \begin{bmatrix} B_d - BD_cD_d \\ -B_cD_d \end{bmatrix}.$$

To detect possible attacks, a parity-space-based residual generator can be constructed. Given order $s \geq n_x$, the parity relation of the desired input $u_s(k)$ and the compromised output $y^a_s(k)$ is expressed as

$$y^a_s(k) = \Gamma_s x(k-s) + H_{u,s} u_s(k) + H_{d,s} d_s(k) + H_{a,s} a_s(k), \tag{5}$$

where $\Gamma_s = \begin{bmatrix} C^\top & (CA)^\top & \dots & \left(CA^{s-1}\right)^\top \end{bmatrix}^\top$ is the extended observability matrix, and

$$H_{u,s} = \begin{bmatrix} D & 0 & 0 & \cdots & 0 \\ CB & D & 0 & \cdots & 0 \\ CAB & CB & D & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ CA^{s-1}B & \cdots & CAB & CB & D \end{bmatrix},$$

is the Toeplitz matrix. Then $H_{d,s}$ and $H_{a,s}$ can be constructed in a similar form of $H_{u,s}$ with $\{B, D\}$ replaced by $\{B_d, D_d\}$ and $\{BB_a, D_a\}$, respectively. As a result, the residual signal $r(k) \in \mathbb{R}^{n_r}$ can be generated as follows [23]:

$$\begin{aligned} r(k) &= P \cdot \Gamma_s^\perp \left[y_s^a(k) - H_{u,s} u_s(k) \right] \\ &= P[W_d d_s(k) + W_a a_s(k)] \end{aligned} \tag{6}$$

where the disturbance and attack projection matrices $W_d = \Gamma_s^\perp H_{d,s}$, $w_a = \Gamma_s^\perp H_{a,s}$, $\Gamma_s^\perp \in \mathbb{R}^{(sn_y - n_x) \times sn_y}$ is the orthogonal complement of Γ, thereby eliminating the impact of the initial state $x(k-s)$ on $r(k)$, and P is the projection matrix with some design freedom. It is obvious that $r(k)$ depends affinely on $d_s(k)$ and $a_s(k)$, and the choice of P shall balance between insensitivity to $d_s(k)$ and sensitivity to $a_s(k)$. It is essential to ensure that the design matrix P has full column rank, avoiding being blind to some not strictly stealthy attack directions $a_s \notin \ker(W_a)$. Subsequently, one can evaluate the residual through the function $J(r) = \|r(k)\|^2$ to detect the occurrence of attacks. The alarm logic is expressed as:

$$\begin{cases} J(r) > J_{\text{th}} \Rightarrow & \text{Attack alarm} \\ J(r) \le J_{\text{th}} \Rightarrow & \text{No alarm} \end{cases}$$

Without loss of generality, the threshold can be set to $J_{\text{th}} = 1$, since its design freedom can be accounted for by optimizing the projection matrix P. Due to the randomness in $d_s(k)$, unwanted false alarms may be raised under the attack-free condition, and FAR has been widely adopted to evaluate robustness performance of the detector. For notational brevity of probabilistic analysis, we use $\xi \in \mathbb{R}^{n_\xi}$ to denote the random variable of the s-long augmented disturbance d_s where the time index k is dropped.

Definition 1 (FAR [5])**.** *Given the threshold $J_{\text{th}} = 1$ and the evaluation function $J(r) = \|r\|^2$, the FAR is defined as*

$$\begin{aligned} \text{FAR} &= \mathbb{P}_\xi \left\{ \|r\|^2 > 1 \mid a_s = 0 \right\} &(7)\\ &= \mathbb{P} \left\{ \|PW_d\xi\|^2 > 1 \right\}. &(8) \end{aligned}$$

To compute FAR, the exact probability distribution of ξ has to be known. The Gaussian assumption, i.e. $\xi \sim \mathcal{N}(0, \Sigma_0)$, is mostly adopted in current literature. As an example, in the GLRT approach the design of P is formulated as a hypothesis test with the null hypothesis $\mathcal{H}_0$ and the alternative hypothesis $\mathcal{H}_1$ [17]:

$$\begin{cases} \mathcal{H}_0 : r(k) \sim \mathcal{N}(0, P\bar{\Sigma}_0 P^\top) \\ \mathcal{H}_1 : r(k) \sim \mathcal{N}(PW_a a_s, P\bar{\Sigma}_0 P^\top) \end{cases}$$

where $\bar{\Sigma}_0 = W_d \Sigma_0 W_d^\top$. Given a preset FAR upper bound α, GLRT gives rise to the threshold $J_{\text{th}} = 1$ and residual generator (6) with the optimal design matrix

$$P = \bar{\Sigma}_0^{-\frac{1}{2}} W_a (W_a^\top \bar{\Sigma}_0^{-1} W_a)^\dagger W_a^\top \bar{\Sigma}_0^{-1} / \sqrt{\chi^2_m(1-\alpha)}. \tag{9}$$

B. Wasserstein-based Distributionally Robust Detection Design

To effectively detect cyber-attacks, a critical issue is how to describe the statistical properties of underlying disturbances are modeled and analyze its impact on detection performance. Under the Gaussian assumption, the well-established GLRT has been widely used to design cyber-attack detectors. However, the true disturbance distribution $\mathbb{P}_\xi$ is typically unknown and show complicated characteristics, e.g. nonstationarity, heavy tails or multimodality, thereby deviating greatly from Gaussianity. When such distributional mismatch occurs, alarm floods can be induced, which make practitioners eventually discredit the attack detector.

To address uncertainty in probability distributions, we draw ideas from DRO to formulate the detector design problem, which captures the unknown disturbance distribution $\mathbb{P}_\xi$ by constructing a so-called ambiguity set $\mathcal{D}$ instead of using a single parametric distribution. Specifically, as long as historical data are available for defining an empirical distribution, one can construct $\mathcal{D}$ as a family of probability distributions that are close to the empirical distribution at hand. To evaluate the distance between two probability distributions, the Wasserstein distance has been a popular option due to its clear interpretability and favorable statistical properties [20], [24].

Definition 2 (Wasserstein distance, [25])**.** *For given two distributions $\mathbb{P}, \mathbb{P}' \in \mathcal{M}(\Xi)$, where $\mathcal{M}(\Xi)$ is the probability space, whose random ξ is supported on Ξ*

$$d_{\text{W}}(\mathbb{P}, \mathbb{P}') = \inf_{\mathbb{Q} \in \mathcal{Q}(\mathbb{P}, \mathbb{P}')} \mathbb{E}_{\mathbb{Q}} \left\{ \|\xi - \xi'\| \right\},$$

where $\mathcal{Q}(\mathbb{P}, \mathbb{P}') \in \mathcal{M}(\Xi^2)$ is the set composed of all the joint distributions of ξ and ξ' with marginal distributions $\mathbb{P}$ and $\mathbb{P}'$.

Definition 3 (Wasserstein ambiguity set, [24])**.** *Given N independent samples $\{\hat{\xi}_i\}_{i=1}^N$, the Wasserstein ambiguity set is defined as:*

$$\mathcal{D}_{\text{W}}(\theta; N) = \left\{ \mathbb{P} \in \mathcal{M}(\Xi) \,\middle|\, d_{\text{W}}(\mathbb{P}, \hat{\mathbb{P}}_N) \le \theta \right\},$$

where $\mathcal{M}(\Xi)$ is the probability space supported on Ξ, θ is the radius of the ambiguity set, and $\hat{\mathbb{P}}_N = \frac{1}{N} \sum_{i=1}^N \delta_{\hat{\xi}_i}$ is the empirical distribution.

Departing from the usual Gaussian assumption, the Wasserstein ambiguity set offers a general data-driven tool for characterizing uncertainty in the true distribution $\mathbb{P}_\xi$ provided that the radius θ is suitably chosen. Following the spirit of DRO, a distributionally robust anomaly detector design problem can be formulated as follows [20]:

$$\max_{P \succ 0} \ \text{tr}\{W_a \bar{P} \bar{W}_a\} \tag{10a}$$

$$\text{s.t.} \ \sup_{\mathbb{P}_\xi \in \mathcal{D}_{\text{W}}} \mathbb{P}_\xi \{\xi^\top W_d \bar{P} \bar{W}_d \xi > 1\} \le \alpha \tag{10b}$$

where $\bar{P} = P^\top P \succ 0$ is introduced to make the objective linear in $\bar{P}$. It maximizes the overall detectability $\|PW_a\|_F^2$ in (10a) while respecting the distributionally robust chance constraint (DRCC) on FAR in (10b).

III. Main Result

A. Stealthy Attack Set and Performance Degradation Metric

The distributionally robust detector (10) only focuses on maximizing the highest overall detectability, while ignoring the adversarial nature of cyber-attacks. The attacker aims to cause damage to the CPS while avoiding raising security alarms. In this work, the attacker is assumed to have access to full system knowledge. Admittedly, for any detector design matrix P, there always exist stealthy attacks that are too unobvious to be detected. This leads to a high security risk that a well-designed stealthy attack may not trigger alarms but do harm to the CPS. Thus we seek to design a detector oriented towards minimizing performance degradation of the CPS subject to FDI attacks. Based on the residual generator and the threshold $J_{\rm th} = 1$, we first define the stealthy attack set $\mathcal{S}_a$ as:

$$\mathcal{S}_a = \left\{ a_s \,\middle|\, \|P(W_d\xi + W_a a_s)\|^2 \le 1 \right\}. \tag{11}$$

We assume that the attack projection matrix W_a has full row rank, such that $W_a W_a^\dagger = I$, and define a new attack representation $\bar{a}_s$:

$$a_s = -W_a^\dagger W_d \xi + \bar{a}_s, \tag{12}$$

By substituting (12) into (11), one can decouple $\mathcal{S}_a$ from ξ, leading to the residual dynamic $r = PW_a\bar{a}_s$. On this basis, we can define an equivalent, but now deterministic stealthy attack set $\mathcal{S}_{\bar{a}}$ using $\bar{a}_s$:

$$\mathcal{S}_{\bar{a}} = \left\{ \bar{a}_s \,\middle|\, \|PW_a\bar{a}_s\|^2 \le 1 \right\}.$$

However, the primary consequence of W_a being row-rank-deficient is the existence of a non-trivial kernel space $\ker(W_a)$. Under this case, attackers can design strictly stealthy attacks $a_s \in \ker(W_a)$, which satisfy $W_a a_s = 0$ and thus will not be reflected in r based on (6). Therefore, we decompose $\bar{a}_s$ into detectable components $\bar{a}_s^{\rm img}$ and strictly stealthy ones $\bar{a}_s^{\rm ker}$:

$$\begin{aligned} &\bar{a}_s = \bar{V}_a \bar{a}_s^{\rm img} + \bar{V}_a^\perp \bar{a}_s^{\rm ker}, \\ &\bar{V}_a = V_{a,[1:sn_y - n_x]},\ \bar{V}_a^\perp = V_{a,[sn_y - n_x + 1:sn_a]} \end{aligned} \tag{13}$$

where the orthogonal matrix $V_a \in \mathbb{R}^{sn_a \times sn_a}$ comes from the singular value decomposition $W_a = U_a \Lambda V_a^\top$. Note that $\bar{V}_a^\perp \bar{a}_s^{\rm ker}$ is completely undetectable by the detector (6), so we set $\bar{a}_s^{\rm ker} = 0$ and merely focus on the detectable component $\bar{a}_s^{\rm img}$. Consequently, the core of the stealthiness analysis reduces to characterizing the set of attacks $\bar{a}_s^{\rm img}$. By substituting (13) into (6), we arrive at $r = PW_a\bar{V}_a\bar{a}_s^{\rm img}$, resulting in the following stealthy attack set under study:

$$\mathcal{S}_{\bar{a}}^{\rm img} = \left\{ \bar{a}_s^{\rm img} \,\middle|\, \|PW_a\bar{V}_a\bar{a}_s^{\rm img}\|^2 \le 1 \right\}.$$

Next we analyze the effect of stealthy attacks on system states in the absence of the true distribution of ξ. The augmented state $\bar{x}$ in (4) can be decomposed as follows:

$$\begin{cases} \bar{x}(k) = \bar{x}^{\rm nom}(k) + \bar{x}^{\rm dev}(k) \\ \bar{x}^{\rm nom}(k+1) = \bar{A}\bar{x}^{\rm nom}(k) + \bar{B}_r y_{\rm ref}(k) \\ \bar{x}^{\rm dev}(k+1) = \bar{A}\bar{x}^{\rm dev}(k) + \bar{B}_a a(k) + \bar{B}_d d(k) \end{cases} \tag{14}$$

where $\bar{x}^{\rm nom}$ represents the deterministic response to the given reference trajectory $y_{\rm ref}$, and $\bar{x}^{\rm dev}$ captures the state deviation caused by both the attack a and uncertain disturbance d. The reachable set of $\bar{x}^{\rm dev}$ can be useful for evaluating performance degradation under stealthy attack. To align with the s-long stealthy attacks $\bar{a}_s$ in $\mathcal{S}_{\bar{a}}$, we first augment the dynamics (14) of the deviation component over a time horizon of s:

$$\bar{x}_s^{\rm dev}(k+s) = \bar{A}_s \bar{x}_s^{\rm dev}(k) + \bar{B}_{a,s} a_s(k+s-1) + \bar{B}_{d,s}\xi(k+s-1) \tag{15}$$

where

$$\bar{A}_s = \begin{bmatrix} 0 & 0 & \cdots & \bar{A} \\ 0 & 0 & \cdots & \bar{A}^2 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \bar{A}^s \end{bmatrix},$$

$$\bar{B}_{a,s} = \begin{bmatrix} \bar{B}_a & 0 & \cdots & 0 \\ \bar{A}_s\bar{B}_a & \bar{B}_a & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \bar{A}_s^{s-1}\bar{B}_a & \bar{A}_s^{s-2}\bar{B}_a & \cdots & \bar{B}_a \end{bmatrix},$$

and $\bar{B}_{d,s}$ are defined akin to $\bar{B}_{a,s}$ with $\bar{B}_a$ replaced by $\bar{B}_d$. One can further replace a_s with $\bar{a}_s$ by substituting (12) and (13) into the augmented system (15):

$$\begin{aligned} \bar{x}_s^{\rm dev}(k+s) = \bar{A}_s \bar{x}_s^{\rm dev}(k) &+ \bar{B}_{\bar{a},s}^{\rm img} \bar{a}_s^{\rm img}(k+s-1) \\ &+ \bar{B}_{\bar{a},s}^{\rm ker} \bar{a}_s^{\rm ker}(k+s-1) + \bar{B}_\xi \xi(k+s-1) \end{aligned} \tag{16}$$

where

$$\begin{aligned} &\bar{B}_{\bar{a},s}^{\rm img} = \bar{B}_{a,s}\bar{V}_a,\ \bar{B}_{\bar{a},s}^{\rm ker} = \bar{B}_{a,s}\bar{V}_a^\perp \\ &\bar{B}_\xi = \bar{B}_{d,s} - \bar{B}_{a,s} W_a^\dagger W_d. \end{aligned}$$

However, ξ may have an unbounded support and is governed by an unknown probability distribution. To carry out set-based reachability analysis, we turn to a distributionally robust confidence set $\mathcal{E}_\xi(\beta) = \left\{ \xi \,\middle|\, \|\xi\|_Q^2 \le 1 \right\}$ with $Q \succ 0$ at a high confidence level $\beta > 0$ (e.g. 0.95), which satisfies the following DRCC:

$$\inf_{\mathbb{P}_\xi \in \mathcal{D}_{\rm W}(\theta, N)} \mathbb{P}_\xi \left\{ \xi \in \mathcal{E}_\xi(\beta) \right\} \ge \beta.$$

This DRCC ensures that $\mathcal{E}_\xi$ is a safe β-confidence set for any distribution $\mathbb{P}_\xi$ residing in the Wasserstein ball $\mathcal{D}_{\rm W}(\theta, N)$. On this basis, we define the following distributionally robust β-level reachable set at time k under stealthy attack:

$$\mathcal{R}_{\bar{x}_s^{\rm dev}}(k) = \left\{ \bar{x}_s^{\rm dev}(k) \,\middle|\, \begin{array}{l} (16),\ i \in \mathbb{N}_{1:k}, \\ \|PW_a\bar{V}_a\bar{a}_s^{\rm img}(i)\|^2 \le 1,\ i \in \mathbb{N}_{1:k}, \\ \|\xi(i)\|_Q^2 \le 1,\ i \in \mathbb{N}_{1:k}. \end{array} \right\}$$

which characterizes possible state deviations caused by both stealthy attacks and disturbance in the β-confidence set. Intuitively, a larger size of $\mathcal{R}_{\bar{x}_s^{\rm dev}}(k)$ indicates severer performance

degradation and consequently, a heavier impact of stealthy attacks on CPS security; however, its explicit dependence on k erects obstacles for evaluating the performance degradation. Therefore, we consider its *asymptotic outer approximation* in the form of an ellipsoid irrespective of k:

$$\mathcal{R}_{\bar{x}_s^{\mathrm{dev}}}(k) \subseteq \mathcal{E}^{\infty}_{\bar{x}_s^{\mathrm{dev}}} = \left\{ \bar{x}_s^{\mathrm{dev}} \;\middle|\; \|\bar{x}_s^{\mathrm{dev}}\|_M^2 \leq 1 \right\},\ k \to \infty \quad (17)$$

where $M \succ 0$ decides both the shape and the volume of the ellipsoid $\mathcal{E}^{\infty}_{\bar{x}_s^{\mathrm{dev}}}$. In a nutshell, $\mathcal{E}^{\infty}_{\bar{x}_s^{\mathrm{dev}}}$ provides an approximation of asymptotically reachable set with stochastic disturbance sampled from the β-confidence set $\mathcal{E}_\xi(\beta)$. It is known that $1/\sqrt{\det(M)}$ is proportional to the volume of $\mathcal{E}^{\infty}_{\bar{x}_s^{\mathrm{dev}}}$ [26], which naturally motivates the utilization of $-\log\det(M)$ as a metric for quantifying performance degradation under stealthy attacks and eventually guiding the design of P.

B. Derivation of Asymptotically Outer Approximation

Our objective is to find the shape matrix M that defines the smallest possible invariant ellipsoid $\mathcal{E}^{\infty}_{\bar{x}_s^{\mathrm{dev}}}$. To this aim, we first present preliminary result, which is useful for developing an asymptotic bound on the state evolution under multiple bounded inputs.

Lemma 1. *([13, Lemma 1]). Suppose a given constant $a \in (0,1)$ and a nonnegative function $V(k)$, if $\omega_i(k)^\top \Omega_i \omega_i(k) \leq 1$ with $\Omega_i \succ 0$, $\forall i \in \mathbb{N}_{1:N_a}$ and there exist $a_i \in (0,1)$, $\forall i \in \mathbb{N}_{1:N}$, satisfying $\sum_{i=1}^{N} a_i \geq a$ and*

$$V(k+1) \leq aV(k) + \sum_{i=1}^{N_a} (1-a_i)\omega_i^\top(k) M_i(k) \omega_i(k), \quad (18)$$

then, the bound of $V(k)$ satisfies

$$V(\zeta(k)) \leq a^{k-1}V(1) + \frac{(N_a - a)(1-a^{k-1})}{1-a},$$

and its asymptotic upper bound is given by $\lim_{k\to\infty} V(k) \leq (N_a - a)/(1-a)$.

Now we are in a position to derive the necessary conditions that the matrix M must satisfy to ensure the ellipsoid $\mathcal{E}^{\infty}_{\bar{x}_s^{\mathrm{dev}}}$ covers. Therefore, Theorem 1 is proposed to formalize these conditions.

Theorem 1. *If (17) with M is a valid asymptotic outer approximation of the reachable set $\mathcal{R}_{\bar{x}_s^{\mathrm{dev}}}(k)$, for a given $a \in (0,1)$, then there exist a_1, a_2 and $\bar{M}$ satisfying:*

$$\begin{cases} \bar{M} = \frac{2-a}{1-a} M \succ 0,\ a_1 + a_2 \geq a \\ 0 \leq a_1 \leq 1, 0 \leq a_2 \leq 1 \\ \begin{bmatrix} a\bar{M} & 0 & \bar{A}_s^\top \bar{M} \\ 0 & \bar{\Omega} & \bar{B}_s^\top \bar{M} \\ \bar{M}\bar{A}_s & \bar{M}\bar{B}_s & \bar{M} \end{bmatrix} \succeq 0 \end{cases} \quad (19)$$

where

$$\bar{\Omega} = \mathrm{diag}\left\{ (1-a_2)\bar{V}_a^\top W_a^\top \bar{P} W_a \bar{V}_a, (1-a_3)Q \right\},$$
$$\bar{B}_s = \begin{bmatrix} \bar{B}_{\bar{a},s}^{\mathrm{img}} & \bar{B}_\xi \end{bmatrix}.$$

Proof. We first define $\zeta(k) = \bar{x}_s^{\mathrm{dev}}(sk)$, $V(k) = \zeta^\top(k)\bar{M}\zeta(k)$, $\omega_1(k) = \bar{a}_s^{\mathrm{img}}(s(k+1)-1)$, $\omega_2(k) = \xi(s(k+1)-1)$, $\omega(k) = \begin{bmatrix} \omega_1(k) & \omega_2(k) \end{bmatrix}$ and $N_a = 2$. By invoking Lemma 1, there exists $a_1 + a_2 \geq a \in (0,1)$, such that (18) holds for any $\bar{x}_s^{\mathrm{dev}}(sk)$. By substituting (16) into (18), we obtain

$$\begin{aligned} &a\zeta^\top(k)\bar{M}\zeta(k) - \left(\bar{A}_s\zeta(k) + \bar{B}_s\omega(k)\right)^\top \bar{M}\left(\bar{A}_s\zeta(k) + \bar{B}_s\omega(k)\right) \\ &+ (1-a_1)\omega_1^\top(k) W_a^\top \bar{P} W_a \omega_1(k) + (1-a_2)\omega_2^\top(k) Q \omega_2(k) \\ &= \begin{bmatrix} \zeta(k) \\ \omega(k) \end{bmatrix}^\top \begin{bmatrix} a\bar{M} - \bar{A}_s^\top \bar{M}\bar{A}_s & -\bar{A}_s^\top M \bar{B}_s \\ -\bar{B}_s^\top M \bar{A}_s & \bar{\Omega} - B_s^\top M \bar{B}_s \end{bmatrix} \begin{bmatrix} \zeta(k) \\ \omega(k) \end{bmatrix} \geq 0, \end{aligned}$$

which implies

$$\begin{bmatrix} a\bar{M} - \bar{A}_s^\top \bar{M}\bar{A}_s & -\bar{A}_s^\top M \bar{B}_s \\ -\bar{B}_s^\top M \bar{A}_s & \bar{\Omega} - B_s^\top M \bar{B}_s \end{bmatrix} \succeq 0 \quad (20)$$

Because $M \succ 0$, we can apply the Schur complement to (20), which directly yields condition (19). This establishes that for the state sequence $\bar{x}_s^{\mathrm{dev}}(sk)$, the asymptotic upper bound $\lim_{k\to\infty} V(k) \leq \frac{2-a}{1-a}$ holds. Due to $a \in (0,1)$, the influence of any initial states $\{\bar{x}_s^{\mathrm{dev}}(i)\}_{i=0}^{s-1}$ vanishes exponentially over time. Therefore, the asymptotically outer approximation of the reachable set $\left\{ \bar{x}^{\mathrm{dev}} \,\middle|\, \bar{x}^{\mathrm{dev}\top} M \bar{x}^{\mathrm{dev}} \leq 1 \right\}$ can be extended to the entire trajectory $\bar{x}^{\mathrm{dev}}(k)$, thus completing the proof. □

C. Design Problem Formulation and Solution Algorithm

The proposed distributionally robust design of FDI attack detector is formalized as follows:

$$\min_{\substack{\bar{P}\succ 0, Q\succ 0, \\ \bar{M}\succ 0, a_1, a_2}} \quad -\log\det(\bar{M}) \quad (21a)$$

$$\text{s.t.} \quad \sup_{\mathbb{P}_\xi \in \mathcal{D}_{\mathrm{W}}(\theta,N)} \mathbb{P}_\xi \left\{ \|W_d\xi\|_{\bar{P}}^2 > 1 \right\} \leq \alpha \quad (21b)$$

$$\sup_{\mathbb{P}_\xi \in \mathcal{D}_{\mathrm{W}}(\theta,N)} \mathbb{P}_\xi \left\{ \|\xi\|_Q^2 > 1 \right\} \leq 1-\beta \quad (21c)$$

$$\text{Constraint (19)} \quad (21d)$$

In this design, the FAR is guaranteed to be below the tolerance level α for any disturbance distribution $\mathbb{P}_\xi$ within the ambiguity set $\mathcal{D}_{\mathrm{W}}$ by the DRCC (21b). The DRCC (21c) defines a safe β-confidence set for ξ that is useful for deriving the asymptotic outer approximation in (21d). Constraint (21d) derived in Theorem 1 is the core of defining the proposed security metric by finding an ellipsoid parameterized by $\bar{M}$ being a valid outer approximation of the asymptotic reachability set of state deviations. Note that minimizing the proposed metric $-\log\det(M)$ is equivalent to minimizing $-\log\det(\bar{M})$. Thus we adopt the convex objective function $-\mathrm{logdet}(\bar{M})$ as the objective, which seeks to reduce the CPS state deviation caused by stealthy attacks as much as possible. However, the optimization problems (21b) and (21c) contain semi-infinite optimization problems, which are computationally intractable. To handle two Wasserstein-based DRCCs, we introduce the following result.

Lemma 2. *([20, Theorem 5]). For given probability threshold α the Wasserstein-based ambiguity set $\mathcal{D}_{\mathrm{W}}(\theta, N)$, the worst-case FAR chance-constraint*

$$\sup_{\mathbb{P}_\xi \in \mathcal{D}_{\mathrm{W}}(\theta,N)} \mathbb{P}\left\{ \|W_d\xi\|_{\bar{P}}^2 > 1 \right\} \leq \alpha$$

holds if and only if

$$\begin{cases} \lambda \geq 0 \\ y_i \geq 0,\ \tau_i \geq 0,\ y_i \geq \lambda - t_i,\ i \in \mathbb{N}_{1:N} \\ \theta + \frac{1}{N}\sum_{i=1}^{N} y_i \leq \lambda\alpha,\ i \in \mathbb{N}_{1:N} \\ \begin{bmatrix} I & -\hat{\xi}_i \\ -\hat{\xi}_i^\top & \hat{\xi}_i^\top\hat{\xi}_i - q_i \end{bmatrix} - \tau_i \begin{bmatrix} W_d^\top \bar{P} W_d & 0 \\ 0 & -1 \end{bmatrix} \succeq 0, \quad i \in \mathbb{N}_{1:N}, \\ \begin{bmatrix} q_i & t_i \\ t_i & 1 \end{bmatrix} \succeq 0,\ i \in \mathbb{N}_{1:N}. \end{cases} \tag{22}$$

In the light of Lemma 2, we can now address the semi-infinite constraints (21b) and (21c) in our design problem. Therefore, Theorem (2) is proposed by replacing them with a finite set of constraints.

Theorem 2. *The distributionally robust attack detector design problem* (21) *admits the following exact reformulation:*

$$\begin{aligned} \min_{\substack{\bar{P},\bar{M},Q,a_1,a_2 \\ \gamma,\lambda,y_i,v_i,\tau_i, \\ \pi_i,t_i,u_i,p_i,q_i}} \quad & -\log\det(\bar{M}) \\ \text{s.t.} \quad & \gamma \geq 0,\ u_i \geq 0,\ \pi_i \geq 0,\ i \in \mathbb{N}_{1:N} \\ & \theta + \frac{1}{N}\sum_{i=1}^{N} u_i \leq (1-\beta)\gamma \\ & \begin{bmatrix} I & -\hat{\xi}_i \\ -\hat{\xi}_i^\top & \hat{\xi}_i^\top\hat{\xi}_i - p_i \end{bmatrix} - \pi_i \begin{bmatrix} Q & 0 \\ 0 & -1 \end{bmatrix} \succeq 0, \quad i \in \mathbb{N}_{1:N} \\ & \begin{bmatrix} p_i & u_i \\ u_i & 1 \end{bmatrix} \succeq 0, i \in \mathbb{N}_{1:N} \\ & \bar{P} \succ 0,\ Q \succ 0 \\ & \text{Constraints (19), (22)} \end{aligned} \tag{23}$$

Proof. The proof is a direct application of Theorem 1 and Lemma 2, so it is omitted for brevity. □

Despite the equivalence between (23) and (21), (23) is a non-convex optimization problem due to the bilinear terms $\tau_i\bar{P}$, $\pi_i Q$, $a_2\bar{V}_a^\top W_a^\top \bar{P} W_a \bar{V}_a$ and $a_2 Q$. Although these BMIs are NP-hard [27], a key observation is that if either set of bilinear variables $\{\tau_i, \pi_i, a_1, a_2\}$ is fixed, the problem becomes a semi-definite program (SDP), which is convex with respect to the other set $\{\bar{P}, Q\}$ and thus becomes solvable. This specific structure motivates us to solve the problem using sequential optimization approach. First, we rewrite (23) as follows:

$$\begin{aligned} \min_{\bar{P}\succ 0, Q\succ 0} \quad & -\log\det(\bar{M}) \\ \text{s.t.} \quad & \mathcal{J}_1(\bar{P}) \leq 0,\ \mathcal{J}_2(Q) \leq 0, \\ & \mathcal{J}_3(\bar{P}, Q) \geq a \end{aligned} \tag{24}$$

where $\mathcal{J}_1(\bar{P})$ is the optimal value of the following subproblem:

$$\begin{aligned} \min_{\lambda,y_i,\tau_i,t_i,q_i} \quad & \theta + \frac{1}{N}\sum_{i=1}^{N} y_i - \lambda\alpha \\ \text{s.t.} \quad & y_i \geq 0,\ y_i \geq \lambda - t_i,\ \tau_i \geq 0,\ i \in \mathbb{N}_{1:N} \\ & \begin{bmatrix} I - \tau_i W_d^\top \bar{P} W_d & -\hat{\xi}_i \\ -\hat{\xi}_i^\top & \hat{\xi}_i^\top\hat{\xi}_i - q_i + \tau_i \end{bmatrix} \succeq 0, \quad i \in \mathbb{N}_{1:N} \\ & \begin{bmatrix} q_i & t_i \\ t_i & 1 \end{bmatrix} \succeq 0,\ i \in \mathbb{N}_{1:N},\ \lambda \geq 0 \end{aligned} \tag{25}$$

$\mathcal{J}_2(Q)$ is the optimal value of the following subproblem:

$$\begin{aligned} \min_{\gamma,v_i,\pi,u_i,p_i} \quad & \theta + \frac{1}{N}\sum_{i=1}^{N} v_i - (1-\beta)\gamma \\ \text{s.t.} \quad & v_i \geq 0,\ v_i \geq \eta - u_i,\ \pi_i \geq 0,\ i \in \mathbb{N}_{1:N} \\ & \begin{bmatrix} I - \pi_i Q & -\hat{\xi}_i \\ -\hat{\xi}_i^\top & \hat{\xi}_i^\top\hat{\xi}_i - p_i + \pi_i \end{bmatrix} \succeq 0,\ i \in \mathbb{N}_{1:N} \\ & \begin{bmatrix} p_i & u_i \\ u_i & 1 \end{bmatrix} \succeq 0,\ i \in \mathbb{N}_{1:N},\ \gamma \geq 0 \end{aligned} \tag{26}$$

and $\mathcal{J}_3(\bar{P}, Q)$ is the optimal value of the following subproblem:

$$\begin{aligned} \max_{a_1,a_2,\bar{M}} \quad & a_1 + a_2 \\ \text{s.t.} \quad & 0 \leq a_1 \leq 1,\ 0 \leq a_2 \leq 1,\ \bar{M} \succ 0 \\ & \begin{bmatrix} a\bar{M} & 0 & \bar{A}_s^\top \bar{M} \\ 0 & \bar{\Omega} & \bar{B}_s^\top \bar{M} \\ \bar{M}\bar{A}_s & \bar{M}\bar{B}_s & \bar{M} \end{bmatrix} \succeq 0 \\ & \bar{\Omega} = \operatorname{diag}\left\{(1-a_1)\bar{V}_a^\top W_a^\top \bar{P} W_a \bar{V}_a, (1-a_2)Q\right\} \end{aligned} \tag{27}$$

This decomposition allows us to perform sequential optimization to resolve the original non-convex problem (23). First, we initialize $\bar{P}$ and Q with small positive-definite matrices to ensure initial feasibility. Then, we can alternately solve three SDPs (25), (26) and (27) using off-the-shelf solvers with fixed $\{\bar{P}, Q\}$ to enlarge the feasible region, and solve the SDP (24) with fixed $\{\tau_i, \pi_i, a_2, a_3\}$ to achieve a higher objective. This procedure is repeated until convergence or the maximum number of iterations. As this iterative scheme is a form of coordinate descent, the objective function is guaranteed to be non-decreasing at each iteration and finally converges. The full implementation details are summarized in Algorithm 1.

Remark 1. *A practical simplification to the design problem* (21) *can be made by coupling the DRCC on FAR* (21b) *and the DRCC on disturbance confidence set* (21b)*. That is, letting $\beta = 1 - \alpha > 0$, then we can choose the disturbance confidence set in asymptotic reachability analysis to be the same set of disturbances that do not trigger a false alarm, i.e. $\mathcal{E}_\xi(\beta) = \left\{\xi \,\middle|\, \|W_d\xi\|_{\bar{P}}^2 \leq 1\right\}$. This strategy can effectively eliminate the need to solve for the matrix Q and the associated BMIs. Consequently, the subproblem* (26) *vanishes in each iteration, thereby improving the computational efficiency of Algorithm 1.*

Algorithm 1 Solution algorithm for reformulation of the distributionally robust detector design problem (23)

Require: Coefficient matrices W_a, W_d, $\bar{A}_s$, $\bar{B}_{\bar{a},s}$, $\bar{B}_\xi$, disturbance samples $\{\hat{\xi}_i\}_{i=1}^N$, Wasserstein radius $\theta > 0$, tolerable FAR $0 < \alpha < 1$ and asymptotic convergence rate $0 < a < 1$.

1: Initialize $\bar{P}^{(0)} \leftarrow \epsilon I$ and $Q \leftarrow \epsilon I$ with a small enough $\epsilon > 0$ and $n_{\rm ite} \leftarrow 0$.
2: **while** stopping criteria not met **do**
3: Solve (25) over $\{\lambda, y_i, \tau_i, t_i, q_i\}$ with $\bar{P} \leftarrow \bar{P}^{(n_{\rm ite})}$, and obtain the optimal solution $\tau_i^{(n_{\rm ite})}$.
4: Solve (26) over $\{\gamma, v_i, \pi, u_i, p_i\}$ with $Q \leftarrow Q^{(n_{\rm ite})}$, and obtain the optimal solution $\pi_i^{(n_{\rm ite})}$.
5: Solve (27) over $\{a_1, a_2, \bar{M}\}$ with $\bar{P} \leftarrow \bar{P}^{(n_{\rm ite})}$ and $Q \leftarrow Q^{(n_{\rm ite})}$, and obtain the optimal solution $a_1^{(n_{\rm ite})}$ and $a_2^{(n_{\rm ite})}$.
6: Solve (24) with $\tau_i \leftarrow \tau_i^{(\rm ite)}$, $\pi_i \leftarrow \pi_i^{(\rm ite)}$, $a_1 \leftarrow a_1^{(\rm ite)}$ and $a_2 \leftarrow a_2^{(\rm ite)}$, and obtain the optimal solution $\bar{P}^{(n_{\rm ite}+1)}$ and $Q^{(n_{\rm ite}+1)}$.
7: $n_{\rm ite} \leftarrow n_{\rm ite} + 1$.
8: **end while**
9: **Return** $\bar{P}$.

Remark 2. *The proposed design involves three key hyperparameters, $\{\alpha, \theta, a\}$, whose selection requires practical consideration. The FAR tolerance α encodes a direct trade-off between detection sensitivity and alarm reliability, and should be set based on the specific application's operational requirements. The Wasserstein radius θ, which governs the level of distributional robustness, can be effectively calibrated using K-fold cross-validation [20]–[22]. Finally, a is the asymptotic convergence rate in the derivation of the invariant outer approximation, which directly influences the shape and conservatism of $\mathcal{E}^{\infty}_{\bar{x}_s^{\rm dev}}$. Since it confines to $(0, 1)$, a can be easily tuned using grid search.*

D. Design under the Case of Row Rank-Deficient W_a

In practice, the attack projection matrix W_a may not have full row rank due to the sensor redundancy or the inherent actuator structure. This section is therefore organized to address this more general and challenging case. We continue to employ the presentation (12) and (13), but redefine

$$\bar{V}_a = V_{a,[1:\mathrm{rank}(W_a)]},\ \bar{V}_a^{\perp} = V_{a,[\mathrm{rank}(W_a)+1:sn_a]} \tag{28}$$

because W_a is row rank-deficient. By substituting (13) into (6), the residual r can be governed by

$$r = PW_a(I - W_a^{\dagger}W_d)\xi + PW_a\bar{V}_a\bar{a}_s^{\rm img}. \tag{29}$$

Because of $W_a^{\dagger}W_d \neq 0$, (29) only eliminates the component of $W_d\xi$ that lies in the row space of W_a, but it cannot completely remove the influence of ξ. To characterize a deterministic set of stealthy attacks $\bar{a}_s^{\rm img}$, we adopt two distinct design philosophies, leading to two different definitions. First, a conservative stealthy attack set $\bar{\mathcal{S}}^{\rm img}$ is defined as:

$$\bar{\mathcal{S}}^{\rm img}_{\bar{a},1} = \left\{ \bar{a}_s^{\rm img} \,\middle|\, \begin{array}{r} \|(I - W_aW_a^{\dagger})W_d\xi + W_a\bar{V}_a\bar{a}_s^{\rm img}\|_{\bar{P}}^2 \le 1, \\ \forall \xi \in \mathcal{E}_\xi(\eta) \end{array} \right\}, \tag{30}$$

which only includes those attacks that can remain undetectable for every possible realization of disturbance ξ in the confidence set $\mathcal{E}_\xi(\eta)$. From the defender's point of view, this is an optimistic assumption that the attacker will only launch attacks that are guaranteed to be stealthy with a probability of at least η. Next, we develop a more tractable approximation of (30).

Theorem 3. *By choosing $\eta = 1 - \alpha > 0$ and $\mathcal{E}_\xi(\eta) = \{\xi \,|\, \|W_d\xi\|_{\bar{P}}^2 \le 1\}$ being a valid η-confidence set, if*

$$\|W_a\bar{V}_a\bar{a}_s^{\rm img}\|_{\bar{P}}^2 \le \frac{1}{2} - \|I - W_aW_a^{\dagger}\|_2^2 \tag{31}$$

then

$$\|(I - W_aW_a^{\dagger})W_d\xi + W_a\bar{V}_a\bar{a}_s^{\rm img}\|_{\bar{P}}^2 \le 1,\ \forall \xi \in \mathcal{E}_\xi(\eta) \tag{32}$$

Proof. To derive a tractable sufficient condition, we upper bound for the left hand of (32):

$$\begin{aligned} &\|(I - W_aW_a^{\dagger})W_d\xi + W_a\bar{V}_a\bar{a}_s^{\rm img}\|_{\bar{P}}^2 \\ &\le 2\|(I - W_aW_a^{\dagger})W_d\xi\|_{\bar{P}}^2 + 2\|W_a\bar{V}_a\bar{a}_s^{\rm img}\|_{\bar{P}}^2 \\ &\le 2\|I - W_aW_a^{\dagger}\|_2^2\|W_d\xi\|_{\bar{P}}^2 + 2\|W_a\bar{V}_a\bar{a}_s^{\rm img}\|_{\bar{P}}^2 \\ &\le 2\|I - W_aW_a^{\dagger}\|_2^2 + 2\|W_a\bar{V}_a\bar{a}_s^{\rm img}\|_{\bar{P}}^2 \end{aligned} \tag{33}$$

Here, the first inequality follows from the Cauchy-Schwarz inequality, the second inequality is due to the property of the spectral norm of matrix, and the last inequality holds by the definition of $\mathcal{E}_\xi(\eta)$. By inserting (31) into (33), we obtain (32), which completes the proof. □

Remark 3. *It is worth noting the tightness of this sufficient condition* (31)*. A noteworthy special case occurs when the singular value of $I - W_aW_a^{\dagger}$ is large enough to make the right-hand side of $\|I - W_aW_a^{\dagger}\|_2^2 > \frac{1}{2}$, leading to an empty $\bar{\mathcal{S}}^{\rm img}_{\bar{a}}$. The physical interpretation is that there exists a large degree of freedom for the disturbances ξ that are orthogonal to the row space W_a to affect the residual r. Consequently, the worst-case disturbance becomes so influential that the alarm can be triggered under any attacks $a_s^{\rm img}$.*

To alleviate conservatism in $\mathcal{S}^{\rm img}_{\bar{a},1}$, we introduce an alternative stealthy attack set $\mathcal{S}^{\rm img}_{\bar{a},2}$ as:

$$\mathcal{S}^{\rm img}_{\bar{a},2} = \left\{ \bar{a}_s^{\rm img} \,\middle|\, \|W_a\bar{V}_a\bar{a}_s^{\rm img}\|_{\bar{P}}^2 \le 1 \right\},$$

which is a broader and more computationally tractable set that characterizes all attacks that are stealthy in the disturbance-free case. This seemingly simple set also bears a clear probabilistic interpretation. If the disturbance ξ is drawn from any centrally symmetric distribution, any attacks $\bar{a}_s^{\rm img} \in \mathcal{S}^{\rm img}_{\bar{a},2}$ are guaranteed to be opportunistically stealthy with a probability of at least 50%.

We now have obtained the well-defined ellipsoidal bounded sets $\mathcal{S}^{\rm img}_{\bar{a},1}$ and $\mathcal{S}^{\rm img}_{\bar{a},2}$, whose structure is the same as that developed for the case of full row-rank W_a. Therefore, the

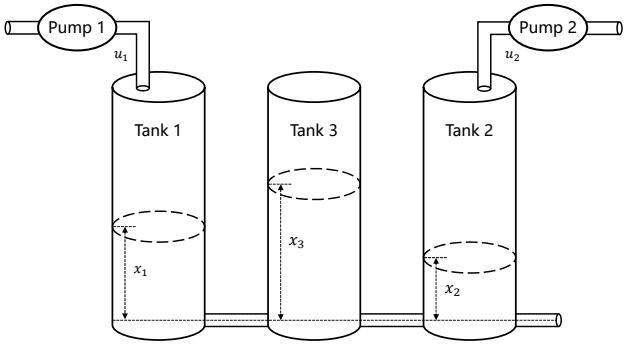

Fig. 2. Schematic of the three-tank system.

corresponding detector design procedure follows a similar spirit to Theorems 1, 2 and Algorithm 1, which is not detailed here since no new ideas are required.

IV. Case Study

In this section, we demonstrate the applicability of the proposed distributionally robust FDI attack detector through a benchmark three-tank system. The laboratory setup known as DTS200 is given in Fig. 2, which consists of three interconnected tank and two pumps. The detailed parameters of DTS200 can be found in [6]. To obtain a discrete-time LTI system, the nonlinear dynamics of DTS200 are linearized around the operating point $x_1 = 10\text{cm}$ and $x_3 = 8\text{cm}$ with sample time $\Delta T = 5\text{s}$, resulting in the system (1) with state-space matrices

$$A = \begin{bmatrix} 0.8586 & 0.0107 & 0.1304 \\ 0.0107 & 0.7958 & 0.1254 \\ 0.1303 & 0.1254 & 0.7390 \end{bmatrix},$$

$$B = \begin{bmatrix} 0.0301 & 0.0001 \\ 0.0001 & 0.0290 \\ 0.0023 & 0.0022 \end{bmatrix},\ C = I_3,\ B_d = I_3,\ D_d = 0.$$

The sensor and actuator attack channels are modeled given by:

$$B_a = \begin{bmatrix} I_2 & 0_{2\times 3} \end{bmatrix}, D_a = \begin{bmatrix} 0_{3\times 2} & I_3 \end{bmatrix}$$

The system is controlled by a feedback controller designed based on an observer and a state-feedback control law.

The controller is composed of a Luenberger observer and a state-feedback control law, resulting in the parameters of (3):

$$A_c = \begin{bmatrix} 0.8078 & 0.0348 & -0.0041 \\ 0.0854 & 0.8017 & -0.0307 \\ -0.0583 & 0.0307 & 0.6471 \end{bmatrix},$$

$$B_c = \begin{bmatrix} -1.5368 & -1.7682 & 0.8636 \\ 0.8590 & 0.9389 & 0.2613 \\ -1.9897 & -0.6146 & -1.5481 \end{bmatrix},$$

$$C_c = \begin{bmatrix} 0.1173 & 0.1297 & -0.2509 \\ -0.3283 & -0.0058 & 0.0179 \end{bmatrix},\ D_c = 0.$$

The residual generator (6) with full row-rank W_a and W_d is obtained by using the parity space method [23] with order $s = 4$ is used to generate residuals. Additive disturbances

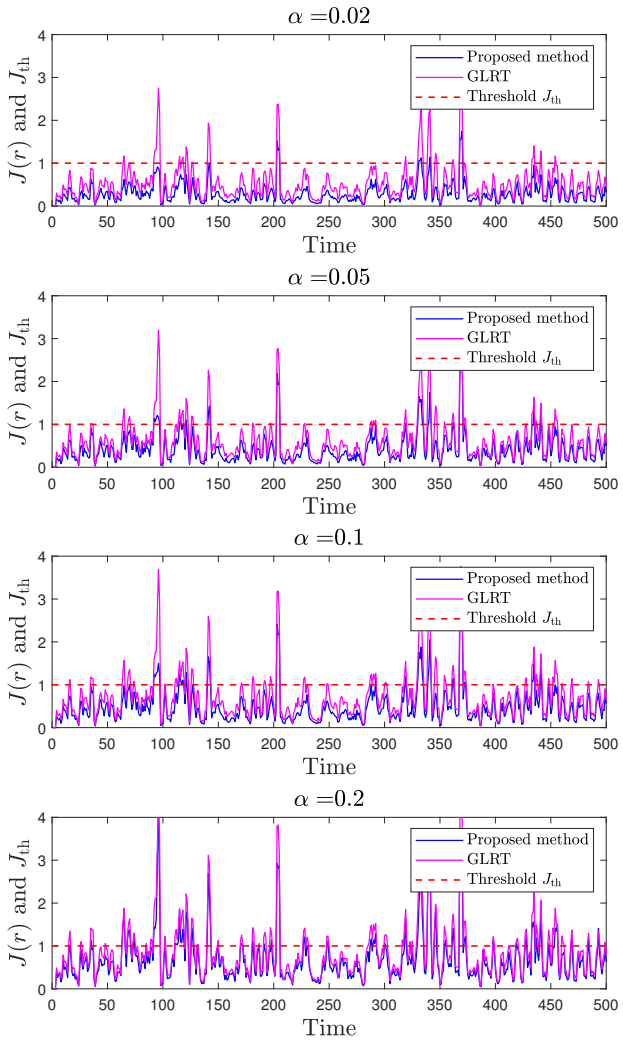

Fig. 3. Residual evaluation function $J(r)$ of the proposed method with $\beta = 0.7$ and the GLRT design under the attack-free case.

$d(k)$ are generated from a Laplace distribution with covariance $\Sigma_d = 0.01I$ and mean $\mu_d = 0$. As for details to implement the proposed method, we use $N = 200$ samples $\{\hat{\xi}_i\}_{i=1}^N$ and the Wasserstein radius $\theta = 0.0029$ to construct the ambiguity set $\mathcal{D}_{\text{W}}(\theta; N)$, and the convergence rate is chosen in the design problem. The induced SDPs (24)-(27) are modeled using YALMIP interface [28] and solved using MOSEK [29] in MATLAB R2024a.

To compare the FARs of our proposed method with the GLRT design (9), we performed Monte Carlo simulations with $200,000$ data points. The FARs of different designs are shown in Table I, and a representative 500 sample is shown in Fig. 3 to showcase the evaluation function $J(r)$ under fault-free cases. As can be seen from the results, the GLRT detector, which is designed under a strict Gaussian assumption, suffers from an "alarm flood" when facing non-Gaussian disturbances. In contrast, the introduction of distributional robustness allows our proposed method to effectively maintain the FAR below the tolerance level α. To visualize the high-dimensional ellipsoid $\mathcal{E}^{\infty}_{\bar{x}^{\text{dev}}}$, we project it onto a lower-dimensional subspace,

TABLE I
FARS OF THE PROPOSED DETECTOR WITH $\beta = 0.8$ AND GLRT DESIGN EVALUATED ON $200,000$ TIME POINTS

	$\alpha = 0.03$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
Proposed method	1.26%	4.85%	7.06%	16.57%
GLRT	9.02%	13.32%	18.35%	26.27%

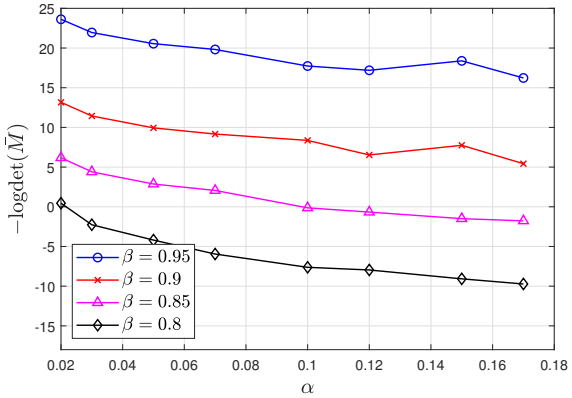

Fig. 4. Residual evaluation function $J(r)$ of the proposed method and the GLRT design under fault-free case.

by partitioning M in (17) as

$$M = \begin{bmatrix} M_1 & M_2 \\ M_2^\top & M_3 \end{bmatrix},$$

where $M_1 \in \mathbb{R}^{n_\mathrm{p} \times n_\mathrm{p}}$. Using the Schur complement, we obtain the following projection $\mathcal{E}'_{\bar{x}^{\mathrm{dev}}}$ on $\mathbb{R}^{n_\mathrm{p}}$:

$$\mathcal{E}'_{\bar{x}^{\mathrm{dev}}} = \left\{ z \in \mathbb{R}^{n_p} \mid z^\top (M_1 - M_2 M_3^{-1} M_2^\top) z \leq 1 \right\}.$$

By setting different values of $\{\alpha, \beta\}$, the performance degradation metric $-\log\det(\bar{M})$ of the induced detector design is displayed in Fig. 4. The induced projection of the outer approximation ellipsoid $\mathcal{E}^{\infty}_{\bar{x}^{\mathrm{dev}}}$ onto the first two state dimensions $\bar{x}^{\mathrm{dev}}_{s,1}$ and $\bar{x}^{\mathrm{dev}}_{s,2}$ by using different α and β is shown in Figs. 5 and 6. As expected, a general trend is observed that the objective value tends to decrease as α increases or β decreases. This can be intuitively explained by the balance trade-off in our design. On one hand, enforcing a stricter FAR constraint, i.e. a smaller α, reduces the sensitivity of the detector, allowing stealthy attacks to cause greater harm and thus resulting in a larger state deviation ellipsoid $\mathcal{E}^{\infty}_{\bar{x}^{\mathrm{dev}}}$. On the other hand, demanding a performance guarantee against a higher confidence set of disturbances, i.e. a larger β, implies that the performance guarantee must hold for a broader set of disturbances, forcing the design to be more conservative, also reflected in a larger $\mathcal{E}^{\infty}_{\bar{x}^{\mathrm{dev}}}$. However, a more nuanced behavior is visible in Fig. 4 that the objective value sometimes does not decrease monotonically with increasing α. This is because our sequential optimization algorithm is guaranteed to converge to a local, rather than global, optimum in the non-convex feasible region of problem (23) with BMIs.

Finally, we depict in Fig. 7 the convergence process of Algorithm 1 under different β, which highlights the effectiveness of Algorithm 1. It can be observed that a larger β leads to

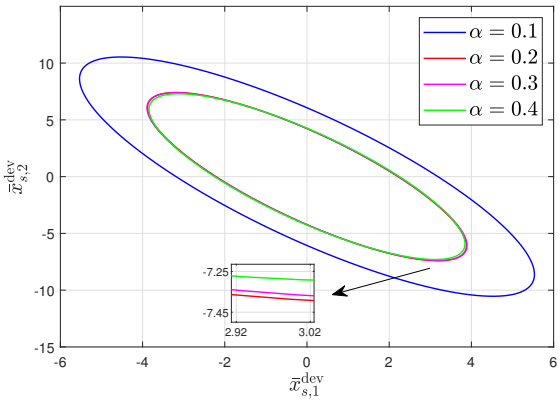

Fig. 5. State deviation reachable set of $\bar{x}^{\mathrm{dev}}_{s,[1:2]}$ versus varying α with $\beta = 0.7$

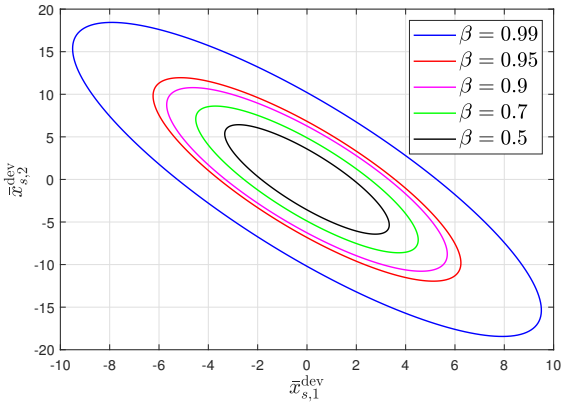

Fig. 6. State deviation reachable set of $\bar{x}^{\mathrm{dev}}_{s,[1:2]}$ versus varying β with $\alpha = 0.05$

faster convergence, and this can be explained by the fact that using a larger β reduces the size of the feasible region of the constraint (20). By operating within a more constrained solution space, the sequential optimization procedure can find a satisfactory solution faster. In contrast, a smaller β enlarges the feasible region, requiring a more extensive search and thus leading to a slower convergence.

V. CONCLUSION

In this work, we proposed a novel detector against FDI attacks in the CPS, with a primary focus on the performance degradation caused by stealthy attacks and stochastic disturbance following an unknown distribution. We first presented a distributionally robust performance degradation metric, which is defined by the volume of an asymptotic outer ellipsoidal approximation of the reachable set of state deviation under stealthy attacks. By optimizing this metric, we formulated the detector design problem while using DRCC to control the FAR under a tolerance level, thereby balancing between FAR and security against stealthy attacks. Next, to address the intractability of the original design problem involving DRCCs, we reformulate it into a finite-dimensional program with BMIs, and devise a tailored solution algorithm based on sequential optimization to efficiently solve the non-convex

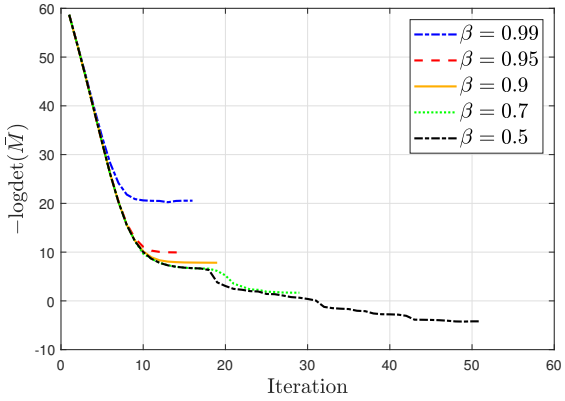

Fig. 7. Iterations of the performance degradation metric in Algorithm 1

problem. Furthermore, we discussed how to extend our detector design scheme to the more general and challenging case of row rank-deficient attack projection matrix, where we formulate the stealthy attack sets in both conservative and optimistic viewpoints. Finally, a case study on a three-tank system demonstrated the efficiency of our design and solution algorithm. For future work, a promising direction is to address strictly stealthy attacks under this distributionally robust paradigm.

REFERENCES

[1] F. Santoso and A. Finn, "A data-driven cyber–physical system using deep-learning convolutional neural networks: Study on false-data injection attacks in an unmanned ground vehicle under fault-tolerant conditions," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 1, pp. 346–356, 2022.

[2] K. Zhang, Y. Shi, S. Karnouskos, T. Sauter, H. Fang, and A. W. Colombo, "Advancements in industrial cyber-physical systems: An overview and perspectives," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 1, pp. 716–729, 2022.

[3] S. N. Edib, Y. Lin, V. M. Vokkarane, F. Qiu, R. Yao, and B. Chen, "Cyber restoration of power systems: Concept and methodology for resilient observability," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 8, pp. 5185–5198, 2023.

[4] Q. Zhang, K. Liu, D. Han, G. Su, and Y. Xia, "Design of stealthy deception attacks with partial system knowledge," *IEEE Transactions on Automatic Control*, vol. 68, no. 2, pp. 1069–1076, 2022.

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

[6] ——, *Data-driven Design of Fault Diagnosis and Fault-Tolerant Control Systems*. Springer, 2014.

[7] Z. Zhao, Y. Huang, Z. Zhen, and Y. Li, "Data-driven false data-injection attack design and detection in cyber-physical systems," *IEEE Transactions on Cybernetics*, vol. 51, no. 12, pp. 6179–6187, 2020.

[8] A.-Y. Lu and G.-H. Yang, "False data injection attacks against state estimation without knowledge of estimators," *IEEE Transactions on Automatic Control*, vol. 67, no. 9, pp. 4529–4540, 2022.

[9] C. Yang, Z. Shi, H. Zhang, J. Wu, and X. Shi, "Multiple attacks detection in cyber-physical systems using random finite set theory," *IEEE Transactions on Cybernetics*, vol. 50, no. 9, pp. 4066–4075, 2019.

[10] K. Liu, H. Zhang, Y. Zhang, and C. Sun, "False data-injection attack detection in cyber–physical systems with unknown parameters: A deep reinforcement learning approach," *IEEE Transactions on Cybernetics*, vol. 53, no. 11, pp. 7115–7125, 2022.

[11] L. Gao, B. Chen, and L. Yu, "Fusion-based FDI attack detection in cyber-physical systems," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 67, no. 8, pp. 1487–1491, 2019.

[12] J.-Y. Ding, K. You, S. Song, and C. Wu, "Likelihood ratio-based scheduler for secure detection in cyber physical systems," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 991–1002, 2017.

[13] C. Murguia, I. Shames, J. Ruths, and D. Nešić, "Security metrics and synthesis of secure control systems," *Automatica*, vol. 115, p. 108757, 2020.

[14] 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.

[15] S. X. Ding and L. Li, "Control performance monitoring and degradation recovery in automatic control systems: A review, some new results, and future perspectives," *Control Engineering Practice*, vol. 111, p. 104790, 2021.

[16] V. Renganathan, N. Hashemi, J. Ruths, and T. H. Summers, "Distributionally robust tuning of anomaly detectors in cyber-physical systems with stealthy attacks," in *2020 American Control Conference (ACC)*. IEEE, 2020, pp. 1247–1252.

[17] A. Willsky and H. Jones, "A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems," *IEEE Transactions on Automatic control*, vol. 21, no. 1, pp. 108–112, 2003.

[18] T. Xue, S. X. Ding, M. Zhong, and D. Zhou, "An integrated design scheme for SKR-based data-driven dynamic fault detection systems," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 10, pp. 6828–6839, 2022.

[19] C. Cheng, Z. Wan, T. Xue, Y. Peng, T. Yin, and H. Chen, "An improved data-driven scheme of robust fault detection for traction drive systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2025.

[20] C. Shang, S. X. Ding, and H. Ye, "Distributionally robust fault detection design and assessment for dynamical systems," *Automatica*, vol. 125, p. 109434, 2021.

[21] C. Shang, H. Ye, D. Huang, and S. X. Ding, "From generalized Gauss bounds to distributionally robust fault detection with unimodality information," *IEEE Transactions on Automatic Control*, vol. 68, no. 9, pp. 5333–5348, 2023.

[22] Y. Feng, H. Jin, S. X. Ding, H. Ye, and C. Shang, "Distributionally robust fault detection trade-off design with prior fault information," *arXiv preprint arXiv:2412.20237*, 2024.

[23] E. Chow and A. Willsky, "Analytical redundancy and the design of robust failure detection systems," *IEEE Transactions on Automatic Control*, vol. 29, no. 7, pp. 603–614, 1984.

[24] P. Mohajerin Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," *Mathematical Programming*, vol. 171, no. 1, pp. 115–166, 2018.

[25] L. V. Kantorovich and S. Rubinshtein, "On a space of totally additive functions," *Vestnik of the St. Petersburg University: Mathematics*, vol. 13, no. 7, pp. 52–59, 1958.

[26] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear matrix inequalities in system and control theory*. SIAM, 1994.

[27] O. Toker and H. Ozbay, "On the np-hardness of solving bilinear matrix inequalities and simultaneous stabilization with static output feedback," in *Proceedings of 1995 American Control Conference-ACC'95*, vol. 4. IEEE, 1995, pp. 2525–2526.

[28] J. Löfberg, "Yalmip : A toolbox for modeling and optimization in matlab," in *In Proceedings of the CACSD Conference*, Taipei, Taiwan, 2004.

[29] M. ApS, *The MOSEK optimization toolbox for MATLAB manual. Version 10.1.*, 2024. [Online]. Available: http://docs.mosek.com/latest/toolbox/index.html